5 Corpora and language engineering

5.1. INTRODUCTION

The aim of this chapter is to provide a *general* overview of the use of corpora in natural language processing (NLP), and to give the reader a *general* understanding of how they have been used and where. This chapter is either a brief excursion for the linguist who is not primarily interested in NLP or a starting point for the linguist interested in corpus-based NLP. In the opinion of the authors it most certainly should not be the end of the road for either type of reader.

In this chapter the focus will be on language engineering. Language engineering is principally concerned with the construction of viable natural language processing systems for a wide range of tasks. It is essentially a 'rather pragmatic approach to computerised language processing' which seeks to bypass the 'current inadequacies of theoretical computation linguistics'.[1] Several areas of language engineering are considered here and the impact of corpora upon them assessed. These areas, split into sections of the chapter, are: part-of-speech analysis, automated lexicography, parsing and multilingual corpus exploitation. Each section considers how the various techniques, employed by specific systems, exploit corpora to achieve some goal. But, before starting to look at the use of corpora in NLP, it seems worthwhile to consider what, in general, corpora can provide for the language engineer.

5.2. WHAT HAVE CORPORA GOT TO OFFER?

Throughout this chapter we will be making frequent reference to a basic distinction that is becoming apparent in language engineering and artificial intelligence, that of cognitively plausible and cognitively implausible systems. Cognitively plausible systems seek to take a model of cognition, thought to be relevant to how humans carry out some task which has intelligence ascribed to it, and to use it as a basis for making a machine carry out that same intelligent task. Such systems often make use of complex sets of rules to express implicit knowledge explicitly, in the form of a **knowledge base**. Systems which eschew cognitive plausibility simply seek to model an intelligent behaviour without making any claims about whether or not the system is operating in the same way as humans do. Such systems often make use of raw quantitative data to generate some statistical model of a behaviour to be mimicked, achieving the same behaviour as a human, but via a clearly different route. This distinction between cognitive plausibility/implausibility is very much a broad-brush one at present and will be refined shortly. In its current form, however, it does allow a general statement to be made about the attraction of corpora for some language engineers.

Corpora can, naturally, contribute equally to the development of cognitively plausible systems as well as to those systems less interested in claims of cognitive plausibility. But it is in the later case, especially where accurate quantitative data are required, that the corpus becomes, by degrees, essential. Quantitative approaches to the solution of problems in artificial intelligence have long had a difficult hurdle to leap, so clearly expressed by McCarthy: 'Where do all the numbers come from?' The answer for NLP at least is now clear – a corpus. As has already been stated in Chapter 1 of this book, humans are poor sources of quantitative data, whereas the corpus is an unparalleled repository of such information. So a good, indeed a balanced, picture of the impact of corpora on language engineering is this: corpora may be used to aid in the construction of systems which are interested in claims of cognitive plausibility – but, where cognitive plausibility is sacrificed to brute force mathematical modelling, corpora are the *sine qua non* of such an approach. Corpora provide the necessary raw data for approaches to language engineering based upon abstract numerical modelling.

It must not be assumed, however, that any system employing statistical modelling is automatically one which eschews cognitive plausibility. We must now refine the broad, polarised description of corpus-based language engineering presented thus far. The sacrifice is almost inevitably a matter of degree rather than an absolute. There are actually very few systems in existence which sacrifice cognitive plausibility *entirely* in favour of, say, a statistical approach. It is far more common for corpus-based quantitative data to be employed as a subpart of a more traditional NLP system based on established methods in artificial intelligence. The task that these data are often most used for is *disambiguation*.[1] In section 5.3.1 on part-of-speech taggers we consider this point in much more detail. Suffice to say, for the moment we will satisfy ourselves with the observation that corpus-based NLP systems which have utterly abandoned standard, pre-corpus, models in language engineering are very few and far between. This chapter actually presents a somewhat skewed view, in that quite a few systems of this sort are reviewed. The reader must bear in mind, however, that these systems have been singled out for presentation here exactly because they are so rare and so radical. As a proportion of all of the systems in language engineering which currently use some stochastic process, they are actually few in number. Most often the corpus and data derived from it are used to enhance the effectiveness of fairly traditional AI (artificial intelligence) systems. The main point is worth reiterating: *any sacrifice of cognitive plausibility is most often one of degree and rarely an absolute.* With this stated we can now begin to consider some specific examples of the use of corpora in language engineering.

5.3. PART-OF-SPEECH ANALYSIS

The development of reliable part-of-speech taggers has had a significant impact upon corpus linguistics. Part-of-speech tagging is now fairly ubiquitous in corpora released in the public domain. Even for corpora built by individual researchers never intended for public consumption, part-of-speech tagging is now achievable to some degree via web-based part-of-speech taggers.[2] In this part of the chapter, we will review how statistically based part-of-speech taggers work. These taggers eschew, to a degree, linguistic rules which may help them in the task of part-of-speech assignment in favour of statistical heuristics which allow tags to be assigned to a word on the basis of what one may describe as contextually aided guesswork. The sacrifice of linguistic knowledge is a matter of degree rather than an absolute, generally speaking. The CLAWS tagger, for example (Garside, Leech and Sampson 1987), did use linguistically motivated rules as a general support to the largely statistical process of part-of-speech tagging in their system. With this stated, let us examine how a tagger such as CLAWS operates.

Assigning parts of speech is an important first step in the automated analysis of a text and is also a task which the corpus linguist may want to undertake in order to aid in the analysis of a text. It is fortunate, then, that part-of-speech information is a useful form of annotation which can be introduced into texts with a high degree of automation (Karlsson *et al.* 1995, Garside and McEnery 1993, Cutting *et al.* 1992, Brill 1992).

The task of part-of-speech assignment consists of assigning a word to its appropriate word class. In the systems developed to date, the traditional basic part-of-speech distinctions, such as adjective, verb, noun, adverb etc., have been supplemented with further relevant information, such as person and number. The size of the categorisation system differs from system to system, though methodologically the systems share some gross affinities, the most notable of which are among the empirically, corpus-based (probabilistic) systems. We will examine in more detail shortly the basic methodology behind such systems as those of Jelinek (1985), Church (1988), Black *et al.* (1993), Déroualt and Merialdo (1986), El-Béze (1993) and Sanchez-Leon and Nieto-Serrano (1995). For the moment, it is sufficient to say that all of these systems share a broadly similar methodology and that this methodology not merely uses, but depends on, the corpus.

This final point is an interesting one. The creation by hand of appropriately annotated corpora can be seen as the raw 'fuel' for some quantitatively based approaches to automated language processing. Again and again throughout this chapter, we will be examining techniques which require annotated text corpora, sometimes of a very large size, in order to be able to go on to automatically generate annotated corpora. There is no chicken-and-egg relation here – before one develops large natural language processing systems to annotate corpora, it seems, at least on current evidence, that one requires a large text corpus, preferably annotated in the form you would like to be able to create, in the first place. So the corpus, unlike the chicken or the egg, definitely comes first! A classic example of this is given in Garside, Leech and Sampson (1987) where the Brown corpus, created by the primitive TAGGIT program of Greene and Rubin (1971) and corrected by hand, was used as the raw data for an empirically based part-of-speech tagger, CLAWS, which went on to out-perform the TAGGIT program, using techniques predicated on the system having access to large bodies of reliable, empirically derived quantitative data, provided in the first instance by the Brown corpus. Other examples will be returned to in later sections. But an important general principle to bear in mind in each section is that corpora are necessary for some modern approaches to NPL.

In our review of part-of-speech tagging, we will limit ourselves to a discussion of statistical part-of-speech taggers. We must accept, however, that purely rule-based taggers have undergone a remarkable renaissance recently. The work of Karlsson *et al.* (1995) in the dependency tagging framework has provided reliable rule-based part-of-speech tagging systems. They have done so largely by developing rules not so dissimilar to those which served the TAGGIT program so poorly. The crucial difference between their work and that of Greene and Rubin (1971) is that modern computers allow for the composition of much more complex rules within which long-distance dependencies between words can be effectively modelled. This was not the case when Greene and Rubin developed the TAGGIT system. The ability to generate such rules has led to the development of rule-based taggers which, arguably, perform better than statistical taggers (Samuelsson and Voutilainen 1997). However, as corpus data had an important historical role to play in the development of statistical language processing systems we will focus on them here. This is not to detract from recent important advances in rule-based part-of-speech tagging, however.

5.3.1. How are corpora used in part-of-speech tagging?

5.3.1.1. How a tagger works

Before we can begin to examine exactly how corpora are used in part-of-speech tagging, it seems useful to describe how part-of-speech taggers work. As noted in the previous section, there are gross affinities between the various systems developed to date. On the basis of these affinities it is possible to posit a general process for the automated part-of-speech tagging of natural language texts (Figure 5.1).

It is possible now to say that the most important stage at which data derived from the corpus intervenes is at the stage of disambiguation. But before

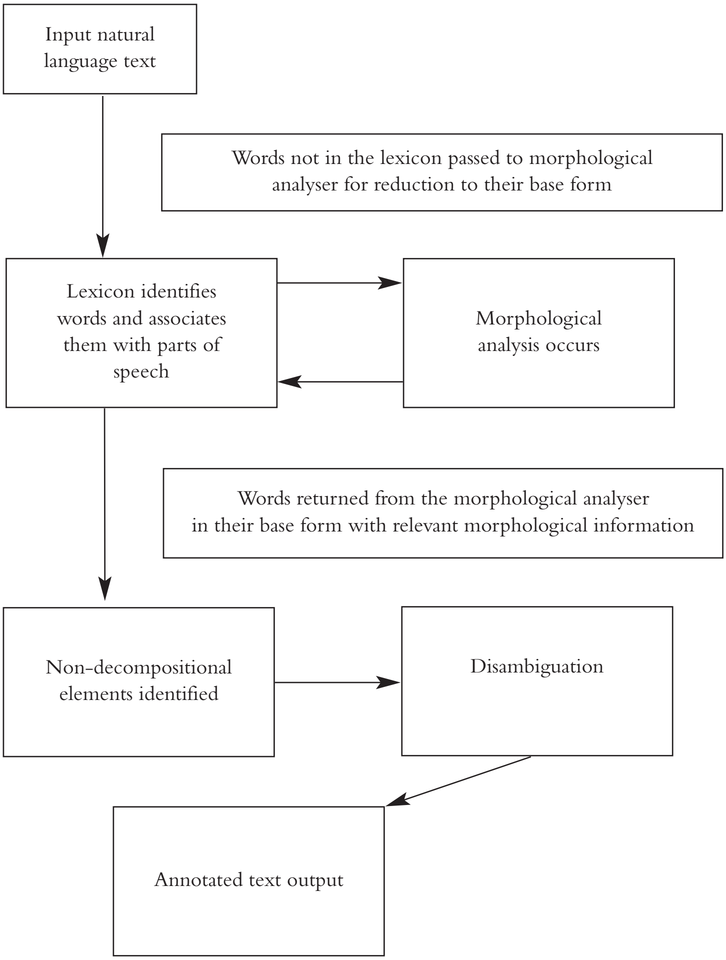
Input natural language text
Words not in the lexicon passed to morphological analyser for reduction to their base form
Lexicon identifies words and associates them with parts of speech
Morphological analysis occurs
Words returned from the morphological analyser in their base form with relevant morphological information
Non-decompositional elements identified
Disambiguation
Annotated text output

Figure 5.1 A schematic design for a part-of-speech tagger

explaining how, let us briefly consider the other stages and consider whether the corpus may have a role to play elsewhere in this schema.

5.3.1.2. Lexicon

The system first tries to see if each word is present in a machine-readable lexicon it has available. These lexicons are typically of the form <word> <part of speech 1, … part of speech n>. If the word is present in the lexicon, then the system assigns to the word the full list of parts of speech it may be associated with. For example, let us say that the system finds the word *dog*. Checking its lexicon, it discovers that the word *dog* is present and that *dog* may be a singular common noun or a verb. Consequently it notes that the word *dog* may have one of these two parts of speech. With this task achieved the lexicon moves on to considering the next word.

Annotated corpora can be of use to this stage of processing. They can, as noted later in this chapter in the section on lexicography, be used as a resource for building lexicons such as these swiftly, automatically and reliably. Without corpora, these resources (often running to hundreds of thousands of entries) have to be constructed by hand. We have only to think back to the limitations of the pseudo-procedure outlined in chapter one to see what a limitation this may be.

Why do such lexicons often run to hundreds of thousands of words? The number of words in a natural language is very large. Some may even argue that there is logically no limit. So the larger the lexicon the better our chances of identifying a word and associating the appropriate parts of speech with it. But what happens when the word in question is not recognised, as it inevitably must be at times? It is at this point that the morphological analyser comes into play.

5.3.1.3. Morphological analysis

Let us assume that the system has now found the word *dogs*. Checking its lexicon, it finds it has no entry for *dogs*. It sends the word to the morphological analyser, which detects an *s* ending and sends the message 'Have found the word *dog* with an *s* ending' back to the lexicon. The lexicon checks to see whether it has the word *dog* in its lexicon. It has, so it reads the tags accordingly – *dogs* can be a plural common noun or a third person singular present tense verb. The morphological analysis is at an end.

Before moving on, however, there are several points to consider. First, note that the morphological analysis carried out is not a true morphological analysis; rather it is guesswork based on common word endings. Some of the word endings that systems such as CLAWS are sensitive to do not even constitute morphemes, for example, -*ly*, which although it often denotes an adjective is not a true morpheme as such.

Second, note that the morphological unit may be called upon several times, as it tries to create different words by removing various endings. Let us assume a rather simple-minded morphological analyser and consider as an example the word *boxes*. Presented with this word by the lexicon, the morphological unit may well see that the word ends with a known common ending, *s*, and send back the message 'Have found the word *boxe* with the ending *s*', only for the lexicon to say that it cannot find the word *boxe*. At this point the morphological analyser may be called upon to try again. Trying the next plausible ending, the morphological analyser sends back the message 'Have found the word *box* with the ending *es*'. At this the lexicon finds the word *box* and proceeds accordingly. The point is that, even with smarter morphological analysers, there is often an interplay between the lexicon and morphology in the automated part-of-speech analysis of a text.

Finally, what if the lexicon, interacting with the morphological component, cannot identify the word at all? Usually the system would resort to contextually aided guesswork – what is the likeliest part of speech for this word given the surrounding words? For the moment we will assume that the system assumes that the word may have any of the accepted parts of speech and leave the assignment of a unique part of speech to the disambiguation phase.

5.3.1.4. Syntactic idioms

What happens where a part of speech does not decompose to the word level, i.e., where two or more words together combine to have one syntactic function, the so called 'syntactic idiom'? Common in this category in English is the multiword preposition (see the section on idiom tags in Chapter 2 for examples). In such cases, part-of-speech analysis systems usually have a specialised lexicon to identify such syntactic idioms and to assign a special set of tags to the words involved in these complex sequences to denote a part-of-speech relation above the level of the word. Corpora may once again be useful sources of information on such sequences and may be used as the basis for an empirically derived specialised lexicon to deal with such features.

5.3.1.5. Disambiguation

After a text has been analysed by the lexicon/morphological processor and any syntactic idioms have been identified, the task of assigning unique part-of-speech codes to words is far from complete. As we saw in earlier examples, the lexicon and morphological component merely indicate the range of parts of speech that may be associated with a word. So, using a previous example, we may know that *dog* may be a noun or verb, but we have no idea which it is in the current context, hence the need for disambiguation. It is possible to try a rule-based approach to disambiguation, using rules created by drawing on a linguist's intuitive knowledge. Indeed, it would be possible to base such a set of rules on a corpus and it would certainly be possible to test such a system on a corpus. But this has not tended to be the use corpora have been put to for disambiguation in part-of-speech tagging. They have, much more frequently, been used to create a matrix of probabilities, showing how likely it is that one part of speech could follow another.

The basic idea behind probability-matrix-based disambiguation is very simple and immediately demonstrable. Consider the following English sentence fragment: 'The dog ate the'. Irrespective of what the actual form of the next word may be, what do you think the part of speech of the next word may be? Most readers would guess that a singular or plural common noun may be a distinct possibility, such as 'The dog ate the bone' or 'The dog ate the bones'. Others may assume an adjective may occur, as in 'The dog ate the juicy bone'. But not many would assume that a preposition, article or verb would occur next. Some may argue that this is because the readers have access to a grammar of the language which shows what is possible and what is not possible.

But this would avoid an important point – why is it that native speakers of a language would feel that some answers are more likely than others? Halliday

(1991) has argued human grammar is essentially a probabilistic grammar. This supports an observation that is very important about language and which was noted by Claude Shannon (Shannon and Weaver 1949) earlier this century. When we utter words, they are not mutually independent in terms of probability. Rather, uttering one word (or class of word) directly influences the likelihood of another word (or class of words) following it. Words are produced not as isolated random events, but as part of a coherent structure. This structure is provided by many features, included amongst which is grammar, and it is precisely structure which is required for a stochastic process such as a transition matrix to operate effectively. The existence of structure in language means that words are linked by dependent probabilities. Part-of-speech taggers such as CLAWS try to use the fact that the structure of language means that words are not mutually independent from one another to disambiguate ambiguous part-of-speech assignments.

Let us return to our example. Given the sentence 'The dog ate the bone', our system may have been able to assign unique part-of-speech tags to all of the words bar *dog*. Here the system would be unsure whether *dog*, in this context, was a noun or a verb. At this point it would consult a probability matrix to decide whether it was more likely that a noun or a verb would be preceded by a definite article and be followed by a verb. On this basis, the system would decide that it was more likely that the word *dog* was a noun rather than a verb. It is in the construction of this matrix of probabilities that the corpus could be most useful. Human beings are relatively poor at assigning a priori probabilities to part-of-speech sequences, or any other language sequences for that matter. However, if we have access to a large, part-of-speech annotated corpus, these probabilities may be reliably and automatically determined. As noted previously, this is exactly the approach described by Garside, Leech and Sampson (1987). Their CLAWS software used a system of probabilistic disambiguation based on probabilities derived from the previously constructed Brown corpus.

To return briefly to a point made previously in the section on morphological analysis, it is also by the use of such a matrix that words unknown to a system are assigned a tag by it. They are assigned the tag that is likeliest given the known surrounding context. Stochastic processes of this sort allow for probabilistic decision making – possibly better described as informed guesswork!

5.3.2. Error rate and processing

Part-of-speech information, as mentioned previously, can now be introduced into corpora with a high degree of automation. The table below shows typical error rates for automated part-of-speech taggers.

With such a high degree of accuracy attainable without human correction, some corpus-building projects are now providing corpora which are wholly automatically annotated with parts of speech. The advantage of this is that processing time and expense can be greatly reduced because human intervention,

Reference	Error rate (%)
Brill (1992)	5
Cutting *et al.* (1992)	4
De Rose (1991)	4
Garside (1987)	4
Greene and Rubin (1971)	23
Voutilainen (1995)	0.7

Table 5.1: Automatic part-of-speech tagging error rates by system

which is slow and expensive, is no longer required. The corollary of this, however, is that any corpus produced by such a method will have a degree of error associated with it, as Table 5.1 clearly shows.

The British National Corpus (Leech, Garside and Bryant 1994) is one such project that has built a corpus using this technique. Leech, Garside and Bryant built a monolingual part-of-speech annotated corpus of 100,000,000 words over a three-year time span. The corpus has a 2–3 per cent degree of part-of-speech classification error associated with it. Another corpus developed with automated part-of-speech analysis was the ITU corpus. Here a parallel aligned English–French corpus was developed with a 2–3 per cent error rate associated with the part-of-speech tagging of each language.

So it is clear to see that part-of-speech taggers, based on quantitative data derived from corpora, exist and are being employed gainfully in corpus-creation projects worldwide. Based upon modest resources initially, these taggers have become a marvellous source of linguistic information and are capable of near-automated creation of part-of-speech annotated corpora.

5.3.3. So what?

Having seen how a part-of-speech tagger may be constructed, it is now possible to begin to see the relationship between corpora and language engineering a little more clearly. Language-engineering systems can be based upon corpus data, which is used to train some model of language which the system possesses. That training can be undertaken on raw – unannotated – corpus data or it may be undertaken on corpus data which has been hand-annotated by linguists. Such hand-annotated data can have at least two functions. Firstly, it may allow the model developed by the program to become more accurate. If the annotation represents reliable evidence of the type of linguistic distinctions that a program wants to model, then the material is likely to provide better training data than a raw corpus.[3] Secondly, such data can be useful as an evaluation testbed for such programs. Rather than relying on human beings to test and assess the output of a part-of-speech tagger, this can be done fairly rapidly and automatically by allowing the part-of-speech tagger to tag text which has already been annotated – if you like, a human-generated crib-sheet is available to the computer to allow it to rate its own performance.

It is clear from this example alone why corpus linguistics and language engineering interact so frequently. Successful language-engineering systems have been developed on the basis of corpus data. Consequently language engineers have called for the creation of corpus data – which is of subsequent use to corpus linguists – and have also developed systems, such as part-of-speech taggers – which are of subsequent use to corpus linguists. In the following section we wish to further this explanation of the relationship between corpus linguistics and language engineering by examining one area where language engineering and corpus linguistics are currently in very close liaison – multilingual corpus linguistics.

5.4. AUTOMATED LEXICOGRAPHY

One way in which lexicography has exploited corpora has been in the creation of a range of automated lexicons. These lexicons in themselves have proved useful in corpus linguistics as a reference resource and as a source of improved corpus-exploitation technology. In this section we will look at the corpus-based construction of:

1. Monolingual lexicons
2. Multilingual lexicons
3. Term banks

Following a brief overview of these topics, a concluding paragraph will consider the impact of corpora on lexicography to date and in the future.

It is worthwhile considering a question before we continue, though: why bother to look in a corpus for lexical information that is available to a trained lexicographer via intuition? Needless to say, the entire discussion undertaken in chapter one addresses this point, and readers may well care to bear in mind this detailed rationale for corpus-based linguistics when considering the reasons why resources developed from corpora are becoming increasingly common; to explain the need for corpus-based lexicography in a nut shell, all corpus-based approaches to lexicography seek to exploit the corpus as a source of information on language that is representative of language, or of a particular genre of language, as a whole, in a way that is economical and reliable.

5.4.1. Automating monolingual lexicon production

Setting aside the practical need for large lexicons if we are to achieve efficient automated part-of-speech annotation, the existence of annotated and un-annotated corpora make the work of the lexicographer somewhat easier and certainly potentially more effective. The application of statistical techniques to monolingual corpora can be an important step in the process of deriving both quantitative and qualitative information for the lexicographer (see Chapter 4). A wealth of frequency based information is available almost instantly at the word level and, if the corpus is annotated with part-of-speech information, a similar wealth of material is available at the syntactic level also.

Qualitatively speaking, the use of measures such as co-occurrence statistics (mutual information, for example, as discussed in Chapter 3) allows us to derive certain semantic relations from the corpus. Lafon (1984) has done work such as this on French, Church and Hanks (1989) on English and Calzolari and Bindi (1990) on Italian. Although the precise method is slightly different in each of these cases, the basic idea remains the same – to postulate semantic affinities between words by measuring the frequency with which words co-occur in close proximity to one another. If words co-occur in a manner which is so frequent that it cannot be ascribed purely and simply to random collocation, then we may assume that there is some special relationship between those words. Note that we are in the realm of dependent probabilities again!

Let us consider the work of Church *et al.* (1991) briefly considered in Chapter 3. They use mutual information to determine a score which expresses the degree to which two words co-occur. Mutual information is essentially an association ratio; how often does a word co-occur with this word as a proportion of the number of times it occurs in the corpus? If the association is one to one, i.e., the words always co-occur, then we may assume that there is some special relationship between the words. Church and Hank's measure assigned a score to each word pairing possible in a corpus (i.e., all possible pairs of tokens from the corpus), assigning values below zero to collocations viewed as random and values above one to collocations viewed as non-random. The further the score below zero, the more likely that a collocation was random. The higher the score above zero, the more likely it was that the collocation was non-random. Their finding was most interesting. Using an association score, two classes of relation were revealed:

1. Between pairs of words bringing together compounds, subparts of set expressions/proper nouns/etc.
2. Between pairs of words with strong semantic associations.

These findings are crucial and are echoed in both the work of Lafon and Calzolari and Bindi. As such, the point seems fairly language independent in a European context and allows for the identification of qualitative data in a corpus using quantitative techniques.

5.4.2. Automating multilingual lexicon production and the construction of termbanks

It is interesting to note that the availability of parallel aligned corpora is making the promise of automated lexicography true in a multilingual as well as a monolingual domain. Using quantitatively based measures, it is possible to extract correspondences between languages not only at the word level, but also above the level of the word. Multiword units may also be retrieved from a parallel aligned corpus, making multilingual dictionary building an easier task. The corpus can also be scanned for frequent collocations.

With a specialised corpus it is possible to construct terminology databases. Such databases are extremely important in machine translation, since within specialised domains there exists (ideally) a one-to-one unambiguous mapping between a concept and its linguistic representation (the term) which makes translation within that domain an easier task (Pearson and Kenny 1991).

Multilingual termbanks are capable of being derived from parallel aligned corpora. As noted by Daille (1994), dynamic and non-dynamic finite state machines can be developed to provide automated terminology extraction, including the extraction of multiword equivalences, where a parallel aligned corpus is available.

5.4.3. Automating lexicon production for part-of-speech tagging

Automated lexicon construction is a goal which seems increasingly realistic thanks to corpus-based NLP systems. Garside and McEnery (1993) and later Smith and McEnery (1997) showed how comprehensive English lexicons for use by part-of-speech taggers can be generated automatically from annotated corpora. The development of such lexicons is not only of use to systems such as CLAWS (Garside 1987): such linguistic resources are prerequisites for bootstrapping models such as the Cutting tagger (Cutting *et al.* 1992). Also, taggers such as the De Rose tagger (De Rose 1991) need the precise word/tag-pair frequencies that the Smith and McEnery (1997) technique produces. Hence the generation of these large part-of-speech tagged lexicons can be seen as an important step towards improving the quality of taggers in any language where appropriate corpora exist.

5.5. CORPORA AND LEXICOGRAPHY

The impact of corpora upon lexicography has been a beneficial one, it would seem (as already noted in Chapter 4). This perceived benefit is one which can easily be seen if we only care to observe the large number of dictionary publishers investing in corpus technology. Browsing along the shelves of any bookshop will add to this impression, as new corpus-based dictionaries will not be hard to find. Corpora have had an important and lasting effect upon work on monolingual lexicography. Major works such as the COBUILD dictionary have been constructed using corpora, and corpora are now being cited by publishers as though they were the very touchstone of credibility for work in modern lexicography. Yet the impact of corpora on multilingual dictionary production has yet to be seen so demonstrably. This is due in part, of course, to the dearth of parallel corpora in comparison to the number of monolingual corpora and to the special needs these corpora present in terms of retrieval. But, as these points are slowly being addressed, it does not take somebody with great powers of prophecy to predict that the changes that have been wrought upon monolingual lexicography will be wrought on multilingual lexicography as soon as parallel corpora become generally available and multilingual retrieval tools more widespread.

5.6. PARSING

As shown in Chapter 4 (section 4.4.4), the field of automated parsing is an active one (see the reference to the Nijmegen work) and a lively one (see the references to the Taylor/Grover/Briscoe–Sampson debate). In some senses, some of the prerequisites for a powerful automated parsing system already exist. To substantiate this claim, consider what such a system should be able to do. It should, minimally, be able to:

1. identify the words of a sentence;
2. assign appropriate syntactic descriptions to those words;
3. group those words into higher-order units (typically phrases and clauses) which identify the main syntactic constituents of a sentence;
4. name those constituents accordingly.[4]

It should also be able to do this with a high degree of accuracy, producing plausible parses for any given input sentence, a goal often described as **robustness**.

Considering the advances made in automated part-of-speech recognition, it is quite easy to see that some of these goals have been as good as achieved. The elusive goal, however, remains the identification of structural units above the word level and below or at the sentence level.[5] Indeed, such is the difficulty of this goal that, if you are reading this book twenty years from its publication date, the authors would not be in the least surprised if no robust parser for general English has yet been created. The current state of the art is somewhat unimpressive for everyday English. Black *et al.* (1993) provide a brief overview of parsing-system testing. The review makes somewhat depressing reading for the ambitious linguist hoping to annotate automatically a corpus with syntactic information above the word level. Satisfactory parses in automated broad-coverage parsing competitions rarely seem to break the 60% barrier and more commonly languish in the region of 30–40 per cent accuracy. To compare this to the state of the art in part-of-speech tagging, recall that the tagger of Greene and Rubin (TAGGIT) achieved an accuracy rate of 77 per cent in the 1970s. Considering such comparisons, we are tempted to conclude, appropriately we think, that the parsing problem is much more complex than the part-of-speech assignment problem. As a consequence the tools currently available are certainly much cruder than part-of-speech taggers and, as a practical tool, probably useless for the corpus linguist at present.

So why discuss it in a book such as this? Well, as with other areas of language engineering , corpora have been utilised to address the parsing problem and again seem to provide the promise of superior performance. To assess the

	Human Rule Creation	**No Human Rule Creation**
Quantitatively motivated	Black *et al.* (1993)	Magerman (1994), Bod (1993)
Non-quantitative motivation	Most traditional AI and language engineering	Brill (1992)

Table 5.2 A summary of different approaches to corpus-based parsing

impact of corpora in the field, it is useful to describe the different approaches that may be taken to the problem by splitting the field into four broad approaches. These are summarised in Table 5.2.

The two axes which differentiate these approaches – Human vs. No human rule creation and Quantitative motivation vs. Non-quantitative motivation – are easily explained. In systems which use human rule creation, there exists within the system a body of rules, such as phrase structure grammar rules, written in some formal notation. These embody certain syntactic principles and describe what the composition of certain constituents may be. Systems using no human rule creation lack this handcrafted set of rules. To move to the second distinction, quantitatively motivated systems attempt to use raw quantitative data from a corpus to perform all or a subpart of the task of parsing. Non-quantitatively motivated systems may use a corpus, but quantitative data derived from the corpus is of no necessary interest to such systems. With this brief explanation of the distinction, it is possible to see why it is useful to quarter the field of parsing system development. In essence, the systems which are quantitatively motivated do not merely exploit corpora as a matter of choice. The corpora are absolutely *necessary* for their operation. Corpora, as noted time and again in this book, are an excellent source of quantitative data, so it is no surprise to see them underpinning a quantitative approach to parsing. Of the non-quantitative approach, we shall see shortly that corpora do have a role to play there also. But it is in the domain of the quantitatively oriented approach to parsing that they reign supreme!

5.6.1. Traditional grammars

Parsing systems in language engineering have most often made use of some linguistic formalism to express syntactic relations. These may be motivated very strongly by theoretical concerns, such as the implementation of parsers based on principles of Government and Binding theory, or may take aspects of syntactic theory, such as phrase structure rules, to encode syntactic relations. Indeed, some AI languages such as PROLOG, have mechanisms which allow the expression of certain aspects of syntactic theory very elegantly. For instance, in PROLOG, the definite clause grammar formalism allows phrase structure grammars to be written in a notation that linguists find familiar and which the machine is able to interpret appropriately.

The variety of systems of this type in existence is so bewilderingly large that it cannot be covered in any real sense here. Suffice to say, these systems are constructed almost exclusively from human created rule sets – large bodies of rules, handcrafted by linguists or system designers, that seek to encode enough human knowledge to construct a credible parser.[6] It will come as no surprise to the reader when it is stated that, except in limited domains, the goal of a robust general English parser by this route seems as distant as when researchers began working towards this goal some three decades ago.

Even if the goal seemed attainable via this route, there is a hideously large problem facing system designers using this route, the so-called 'knowledge acquisition bottleneck'. This is a common problem in AI – how do we sit down and write a huge set of rules which describe how to perform some task accurately and consistently? The problem seems at best unpalatable and at worst impossible. We will return to a brief discussion of this topic in section 5.7 on machine translation later in this chapter. For the moment note that another limitation of this approach is the scale of the task of enumerating a set of rules which would encapsulate a credible grammar of a natural language.

But what of the use of corpora in such systems? Corpora are having an impact in this area, most commonly as training sets for parsers developed by reference to intuition. Some researchers are also trying tentatively to include some corpus-based stochastic modelling within linguistic models of syntactic competence, in order to improve the efficiency of parsing systems based on these. An example of this is the work of Fordham and Croker (1994). These researchers sought to use a corpus to parameter set and control aspects of a Government and Binding parser. Such work appears to be in its infancy, but indicates that a synthesis of probabilistic and cognitively plausible approaches to parsing is an idea that has penetrated the area of computational natural language parsing. As we shall see in section 5.6.3, some researchers have taken this synthesis further. An important distinction to note, however, is that Fordham and Crokers' work is a direct amendment of current linguistic theory, while the work of Black *et al.* (1993) described later is decidedly pre-theoretical. As a consequence, the linguist and especially the psycholinguist may find more attraction in the goal of employing and justifying some stochastic elements in current theories of language, rather than relinquishing the goal of theoretically motivated cognitive plausibility to move to a more pre-theoretic, or even linguistically atheoretic, account of language.

5.6.2. Radical statistical grammars

If traditional AI systems stand at one end of a continuum representing approaches to parsing, then what are described here as radical statistical grammars stand at the opposite extreme. These systems are what were called *linguistically atheoretical accounts* of language in the previous section. They are called that here because they eschew all expert linguistic knowledge expressed as handcrafted rules – the grammars derived by these systems are most certainly not any which a linguist would recognise or that anyone would claim had the remotest claim to cognitive plausibility.

Radical statistical grammars seek to use abstract statistical modelling techniques to describe and ascertain the internal structure of language. At no point is any metaknowledge used beyond the annotated corpus, such as a linguist's or system designer's intuitions about the rules of language. The system merely observes large treebanks and on the basis of these observations decides which structures, in terms of word clusterings, are probable in language and which are not. In many ways it is best described as a statistical derivation of syntactic structure.

Magerman's system is an interesting example of this. Magerman (1994) looks at a sentence in a radical way – he assumes that any 'tree' which can fit the sentence is a possible parse of the sentence. There is no pre-selection of trees – Magerman simply generates all right-branching trees that fit a given sentence. He then clips through this forest of parses and uses his corpus-based probabilities in order to determine what is the most likely parse. Magerman's system uses little or no human input – the only implicit input is that language can be analysed as a right-branching tree and annotated corpora are used as a source of quantitative data. Otherwise linguistic input to Magerman's model is nil. Bod (1993) produces similarly radical work, again organising language independent of any handcrafted linguistic rules.

The work of Magerman and Bod is genuinely radical. It is common to find works which, as this book has noted, merely augment traditional AI systems with some stochastic information and claim that a radical paradigm shift has taken place. An example of this is Charniak (1993), who develops a probabilistic dependency grammar and argues that this constitutes a major paradigm shift. A careful examination of Charniak's work reveals that the major paradigm shift is constituted by the incorporation of statistical elements into a traditional parsing system, not surprisingly, at the stage of disambiguation. When we compare this to the radical paradigm shift Magerman or Bod promise, we see that, while Charniak's work is, in some ways, novel, it is by no means as powerful a radical paradigm shift as that proposed in the work of Magerman and Bod.

The main point of importance for the corpus linguist is this: if the radical statistical paradigm shift in language engineering is to take place, corpora are *required* to enable it. Both Magerman and Bod rely on annotated corpora for training their statistical parsers. Without corpora the a priori probabilities required for a linguistically atheoretical approach to parsing would be difficult to acquire. With this stated, it is now possible to move on and consider the pre-theoretical hybrid approach to parsing mentioned in the previous section.

5.6.3. A hybrid approach – human rules and corpus statistics

This approach was pioneered by Black *et al.* (1993) at IBM. Here the system uses corpus statistics to choose between a variety of conflicting parses available for a sentence produced by a large handcrafted grammar. The parses are of a pre-theoretical nature; although they are useful surface parses of a given sentence, the parsing formalisms are not motivated by a particular linguistic theory and do not produce parses which would conform to all of the needs of, say, Extended Standard theory or Dependency Grammar theory. But they could provide a useful first step to most forms of theoretically motivated linguistic analysis.

Even so, Black *et al.*'s system had to deal with a classic problem for parsing-system designers which has been touched on already in this chapter: ambiguity. One of the recurrent problems for the builders of parsing systems, is that, while there seems little or no difficulty in ascribing structures to sentences, there is a certain overproductivity entailed in the operation – in short, any given sentence may have many potential parses. Which one do you choose and why? Take the sentence *The dog heard the man in the shed*, for example. Whether we believe the man or the dog to be in the shed is one crucial factor in deciding how this sentence is to be represented. This example is based upon a classic problem – prepositional phrase attachment; in essence, in the absence of further evidence, the attachment seems a fairly arbitrary decision in contexts such as this. But, where we have a corpus that is appropriately annotated (a treebank), it may be possible to indicate which rules of attachment are *most likely* to be active in such a syntactic context, by examining the frequency with which these rules were invoked in the corpus in similar contexts. That is precisely what Black *et al.* sought to do. The human-created rules were conditioned with empirically derived probabilities extracted from a treebank. For any given situation covered by the human-created grammar, it would be possible not only to interrogate the rule base to produce a series of parses that were potentially plausible analyses of the sentence, but also to rate the plausibility of those parses by manipulating probabilities associated with the rules invoked to produce them.

Black trained his grammar on a training corpus, using a learning algorithm called the 'inside outside algorithm',[7] to condition his rule base with probabilities. Black did this on a corpus of computer manual texts in the first instance, creating a parser for that genre. He then used a test corpus to assess the accuracy of his system. The test corpus was also a treebank and accuracy was gauged with respect to the decisions made by human analysts.

He found that roughly 1.35 parses per word was the norm for the grammar that he developed. In other words, for a 10-word sentence, the system would have to choose between 13.5 potentially plausible parses on average. Not surprisingly then, the success of Black's system decays as the average length of sentence exposed to the system increases. But even so, the system attained some respectable accuracy scores, with results running from 96 per cent to 94 per cent for computer manual sentences averaging 12 and 15 words respectively. Black's work is interesting as it shows how raw quantitative data and traditional linguists' grammars can be combined to achieve a useful goal – accurate parses of sentences in restricted domains. It also shows again, quite clearly, the need for appropriately annotated corpora in language engineering.

5.6.4. Non-quantitative corpus-based grammars

Brill (1992) produced an interesting parsing system, though it was a pre-theoretical system once again. It was one which required no handcrafted grammar rules, but which also avoided the use of statistical learning and decision making. This final class of grammar parallels the work of Magerman and Bod in some respects, in that it is again based solely upon an annotated corpus – no handcrafted rules are used by the system at all. The grammar is induced directly from the corpus and no human intervention or expert knowledge is needed beyond that encoded in the corpus annotation. As noted, it differs in what it does with the corpus – it does not use the corpus as a source of quantitative data. Rather it uses the corpus annotation to induce a set of rules which allow it to generate parses for a given sentence.

When the system has derived the grammar from the corpus annotation, it can proceed to analyse new sentences. At this point another important difference between the systems of Brill and Magerman becomes apparent. Brill's system does not use probabilities to decide between conflicting parses. Rather it uses a heuristic called 'rules of bracket permutation'. This allows the system to assign a set of default brackets to a sentence and then to shift the brackets until the system is happy with the result. In other words, the system assumes a parse for a sentence in the first instance and then literally plays around with the brackets to see whether it can get a parse that better matches the nature of the sentence, a decision made based upon the system's experience with the annotated corpus. The system adds, deletes and moves brackets until it judges that it has found the optimal parse for the sentence.

The minutiae of how Brill's system works are not of great interest to us here. The main point of interest is that Brill's system not merely has the option to exploit corpora – rather the entire approach is predicated upon access to suitable corpora. More specifically, it is predicated upon access to suitably annotated corpora. Corpora seem once again to be permitting a form of research in language engineering that would otherwise be inconceivable.

So what, in summary, can we say about the impact of corpora upon parsing systems? If the four-part distinction between parsing systems outlined is credible, then we are forced to the conclusion that most paradigms of research into computational parsing require corpora, especially annotated corpora, and without them only traditional broad-coverage parsers of English would be under construction. Although traditional systems could make use of corpora, as training data, for example, or as a crutch to support the linguistic intuition of the system designer/linguist, the approach does not *require* corpora. It is hardly surprising, therefore, to discover that in this field corpora have had relatively little impact. However, beyond the traditional systems, the other three approaches to parsing have been *enabled* by the creation of suitably annotated and reliable corpora. There would only be traditional parsers without corpora. The importance of this observation needs underlining. If we take Black's (1993: v) observation regarding traditional parsers to be true:

> The programs that do exist … are actually not very good even at establishing the event pool for everyday sentences, let alone choosing the intended event [the 'intended parse'].

then, if corpora had never been developed for anything else, people may have had to create them anyway if progress towards the construction of a broad coverage parser of English was to be made. Traditional parsing systems have seemed to show no promise for a long time. Not only were they limited in terms of accuracy, but they also faced another classic AI problem – scaling up. In short, it is relatively easy to build an AI system that acts credibly within some highly constrained subdomain. The problem comes when you try to build any system up to operate in the world at large, not in some falsely circumscribed subdomain. Just as with other AI problems, traditional parsers face a problem of scale. They may work well in a world in which only 100 grammar rules and 1,000 words are allowed. But wishing to expand the system to deal with an infinitely generative set of rules and an ever-expanding lexicon creates a hurdle that all systems seem to fall at. Corpus-based broad-coverage parsing systems are in the ascendant purely and simply because they seem easier to create and they produce better results than traditional systems. Corpora have facilitated positive developments in an area of language engineering that has been unpromising for some time.

5.7. MULTILINGUAL CORPUS EXPLOITATION

We have already mentioned, in section 2.3, at least one important exploitation tool which increases the usefulness of parallel corpora – corpus alignment. Alignment can occur at many levels, but at the moment sentence and word alignment are by far the most common forms of alignment available.[8] In order to exemplify current work in alignment we intend once more to outline work at Lancaster and then to set this work in a broader context by a wider review. Following this we will consider how corpus retrieval tools (concordancers) are being developed to allow humans to exploit multilingual corpora effectively.

5.7.1. Alignment

Alignment at sentence, word and multiword unit level is seen as a key process in the exploitation of parallel corpora for a variety of purposes, including the following:

- Example-based machine translation (Nagao 1984)
- Statistically-based machine translation (Brown *et al.* 1990)
- Cross-lingual information retrieval (Melamed 1996)
- Gisting of World Wide Web pages (Melamed 1996)
- Computer-assisted language learning (Catizone *et al.* 1989, Warwick-Armstrong and Russell 1990)

All of the above applications of parallel corpora are, however, dependent to lesser or greater degrees on appropriate alignment software being available. Currently, alignment proceeds:

- on the basis of a statistical heuristic
- on the basis of linguistic rules
- on a combination of the two above.

Let us briefly consider these three approaches. Statistical alignment techniques employ heuristics empirically to achieve alignment. Often simple quantitative measures can be used to determine sentence-level translation equivalents, given two parallel translations, with good results. For example, the approaches of Brown *et al.* (1991) and Gale and Church (1991) depend on relative sentence lengths, based on the premise that long sentences in one language are more likely to be translations of long sentences in the other, while short sentences in one language are more likely to be translated by short sentences in the other.[9] Melamed (1996) uses techniques originally developed in automated image processing in order to achieve sentence and word alignment.

The alternative approach is to employ linguistically motivated methods. These approaches are rationalistic, often inspired by what we might do intuitively when manually aligning texts. Linguistic methods are generally based on pairing lexical units which make up phrases, eventually accompanied by their dependency structures.

The statistical and linguistic approaches are not mutually exclusive, but are complementary and can be usefully hybridised. Statistical methods tend to work better for large corpora, since they are relatively rapid, while linguistic methods can be better for small corpora (Debili and Sammouda 1992). Further, the two techniques can be combined, for example, Chen (1993) found that the most reliable indicators for alignment were ‘critical parts of speech’, namely, nouns, verbs and adjectives, which could be used to support an otherwise statistical alignment method.

Sentence alignment software has been developed and tested, based upon a variety of techniques as described by Gale and Church (1991), Kay and

Röscheisen (1993) and Garside *et al.* (1994). Such alignment programs allow users to align parallel corpora as a prelude to exploitation in, for example, lexicographic research. Table 5.3 gives a series of results obtained at Lancaster using the Gale and Church technique (which is based upon statistical heuristics) for sentence alignment (taken from McEnery and Oakes 1996).

Word alignment software has been developed as a further aid to corpus exploitation, based, for example, upon approximate string matching techniques and co-occurrence statistics. Table 5.4 gives an example of the results achieved on the English–Spanish language pair using Dice's similarity coefficient (McEnery and Oakes 1995). The Dice score measures the similarity of two words (say, one French, one English). The closer to 1 the Dice score is, the more likely it is that the two words are translations of one another. The closer to 0 the score, the less likely it is that the two words in question are translations.

Multiword unit alignment software has also been developed by a range of researchers. For example, Gaussier *et al.* (1992), worked on English–French multiword unit alignment. Their work was extended to cover the English–Spanish and French–Spanish language pairs by McEnery *et al.* (1996). Some results which illustrate this are given in Figure 5.2 below (taken from McEnery *et al.* 1996). The results are based upon a combination of statistical heuristics and linguistic knowledge. Likely matching sequences are extracted from the corpora using a knowledge of noun-phrase structures between the two languages. A mutual information statistic (see section 3.4.4 for further details of this measure) is then used to test whether the words composing the candidate terms are in fact in frequent association with each other. Figure 5.5 shows how, combining the linguistic and heuristic information, the automated extraction of technical terms from parallel corpus texts may be achieved more accurately. While the overview of alignment techniques presented here is, of necessity,

Language pair	Domain	Paragraphs tested	% correct
English–Polish	Fiction	89	100.0
English–French	Telecommunications	100	98.0
English–Spanish	Telecommunications	222	93.2
English–German	Economics	36	75.0
Chinese–English	Newspaper	171	54.5

Table 5.3 The success of sentence alignment

Dice score	0.4–0.49	0.5–0.59	0.6–0.69	0.7–0.79	0.8–0.89	0.9–0.99	1.0
Accuracy	41.1%	92.7%	90.2%	100%	100%	100%	100%

Table 5.4 The success of word alignment in English–French using Dice's similarity coefficient

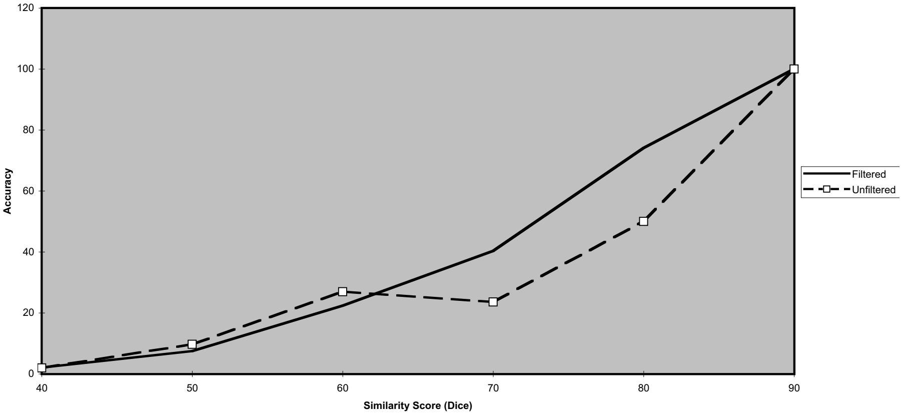

Figure 5.2 The success of compound noun alignment between English and Spanish using finite state automata and similarity data both with and without co-occurrence measures as an additional filter

brief and sketch-like, it is possible to see from the work presented that a range of techniques have been developed to achieve alignment. Further to this, sentence-alignment technology has proved to be fairly reliable on the language pairs it has been tested on so far.While word alignment is still not as reliable, advances have been made towards this goal also, and work on phrasal alignment is under way. Aligned data is clearly of use for machine-translation tasks (see section 5.7.3), but what of machine-aided translation? At least part of the answer to this question lies in the development of parallel concordancers which allow translators to navigate parallel text resources.

5.7.2. Parallel concordancing

To make parallel corpora easy to exploit it is clear that we require a new type of concordancer, one which can deal with parallel aligned corpora. Although one could imagine carrying out two monolingual searches through the L1 and L2 texts of a parallel corpus, it would be of greater advantage to carry out a search in, say, L1 and, as part of the retrieval of relevant context, the program displayed both the L1 contexts and their L2 translations.This would be a way of providing translators with an online translation memory of sorts.

With such a need identified, it is hardly surprising that multilingual concordancers are becoming available. The WordSmith program devised by Mike Scott at the University of Liverpool contains a rudimentary alignment algorithm. More sophisticated is the MultiConc program produced by Woolls (2000) at Birmingham.This uses a modified version of the Gale and Church alignment algorithm to align texts 'on the fly' as they are presented to the system. The system is also capable of working in a wide range of fonts. Less sophisticated is the ParaConc program developed by Barlow (2000) which allows multilingual concordancing but only on texts which have been explicitly pre-encoded with alignment information. ParaConc has no online alignment facility. Of some interest is the growing number of web-based multilingual corpus browsers, such as that of Peters, Picchi and Biagini (2000), which allow remote concordancing of multilingual texts. All of these systems, especially those with online text alignment facilities, represent a possible way of translators exploiting available corpus data and data they themselves generate as a form of translation memory. As translation memories form an important part of the use of parallel corpora in machine translation, we shall now move to an examination of the role of parallel corpora in that field.

5.7.3. Machine translation

Machine translation (MT) systems and research programmes are increasingly making use of corpora, especially parallel aligned corpora. The appeal of corpora to machine translation can be appreciated fairly quickly, if we refer back to the problem of scaling up AI systems as discussed in the last section. If enormous amounts of world knowledge and formally expressed rules are needed to achieve effective automated parsing of language, imagine how much more world knowledge and how many more formal rules are necessary for translating between languages. The size of the problem has been described as being prohibitively large. The systems developed for MT need to routinely build massive resources, yet the technology and techniques for this process do not yet exist. In the field of MT studies several avenues of escape from this problem of scale have become apparent over the past decade or so.

The first, and possibly, we may argue, the most painful avenue is to laboriously construct, by hand, the resources needed for these full-scale systems. Some initiatives are under way in this area, including the Cyc project of Lenat and Guha (1990). This project aims to build a world-knowledge encyclopaedia which machines would be able to use. Yet, to give an idea of exactly how painful this process may be, Lenat has been working for nigh on a decade and has yet to put his results in the public domain. Even when the results are publicly available, it is inevitable, as pointed out by Minsky and Riecken (1994), that the result will be a first approximation to the eventual solution by this route. A satisfactory machine-exploitable world-knowledge base may be a long, long way into the future.

Needless to say, some see ways of accelerating this process and it is at this point that corpora make their entrance into the field of MT studies. Some have sought to extract relevant information from text corpora to speed the process of compiling effective resources for MT systems. Smadja (1991) used text corpora in an attempt to automate the process of knowledge acquisition for MT systems, while McEnery *et al.* (1997) used a trilingual English–French–Spanish corpus and similarity metrics to extract a series of language-specific cognates for this language pair using fuzzy matching techniques. Such systems develop resources for existing MT systems by exploiting corpora as sources of, say in the case of McEnery *et al.*, bilingual lexical knowledge. But some researchers have gone further in their application of text corpora to the MT problem and, in doing so, have developed new paradigms of machine translation. Of these paradigms, two in particular seem predicated upon the use of corpora, especially of the parallel aligned variety. These approaches are statistical translation and example-based machine translation (EBMT).

5.7.3.1. Statistical translation

Statistical translation is very much one of those rare examples of what this chapter has termed genuinely radical quantitatively based computational linguistics. It is quite unlike, say, probabilistic tagging. The approach does not amend an existing model, patching it up with statistics where necessary. Statistical MT entails the total rejection of existing paradigms of MT, based upon fairly traditional AI, in favour of an approach to MT which totally eschews any goal of cognitive plausibility. As a consequence, we see notions of transfer, interlanguages and rule-based grammars, the common staples of most MT, replaced by an approach to translation based upon complex calculations using co-occurence data and distribution probabilities derived from corpora. There is a telling anecdote from the time that the early work on this approach was undertaken on this at the IBM T. J. Watson Research Center. The story is this: the manager in charge of this research noticed that every time he sacked a linguist the system's accuracy increased. Although possibly apocryphal, this story exposes the essential idea behind the system of most radical non-cognitively plausible models in computational linguistics; are models of cognition relevant to human beings the best models on which to base computer models? We shall not seek to discuss this question here. It is simply interesting to note that corpora are fundamental to non-cognitively plausible approaches to language engineering as they actively seek to model performance.

With this stated, let us briefly consider one such system. The first approach to MT taken within this paradigm was that of Brown *et al.* (1990, 1993). This work attempted to build upon the success of probabilistic methods in other areas of language processing and apply them to the problem of machine translation. Brown *et al.* implemented a fully probabilistic machine translation system trained on an aligned French–English corpus. Given a sentence in the source language, this system chose the most probable translation sentence in the target language, using two probability models: a translation model based on lexical alignments and word-position probabilities, and a language model based on trigram probabilities (i.e. sequences of three words).

Brown *et al.* achieved success rates that seemed impressive, but a key criticism of their work stands: their work dealt only with English and French. Would this method work on languages which differed more, syntactically and lexically? For the moment one can only speculate as to what the answer to this question would be. Future work depends upon the creation of corpora aligning language pairs other than English and French. Such work is under way and it could be hoped that, some time soon, there will be an answer to this criticism. For the moment we must keep an open mind. But, the fact that we must wait, admirably underscores the point that this approach to MT requires parallel aligned corpora. Corpora are not an option here. They are a necessity.

5.7.3.2. Example-based machine translation

The first approach to EBMT was initially proposed by Nagao (1984) and elaborated at the sub-sentence level by Sadler (1989). Nagao had a basic, yet elegant, idea. If we have two corpora, aligned at the sentence level, say in language A and language B, then, if we are asked to translate sentence *x* from language A into some sequence in language B, we can consult our corpus to see if we already have sentence *x* in language A. If we do indeed have this sentence in our corpus, then we may find its translation, via alignment, in language B. We have the translation of *x* without having done anything beyond examining our parallel aligned corpus.

The basic EBMT idea has proved very powerful in language engineering. The prospect of relatively painless translation systems has led to a growing amount of work along these lines. Yet the limitations of the basic EMBT idea are obvious. As noted in Chapter 1 of this book, the prospect of the corpus being the sole explicandum of language via enumeration is one which is not credible. It therefore stands that any given parallel aligned corpus, no matter how large it is, can only cover a subset of all of the possible required translations between two unconstrained natural languages at the sentence level. Sadler (1989) built upon the basic idea of EBMT and suggested a more elaborate method of exploiting parallel corpora for MT systems. This approach entailed a very large bilingual database being constructed, with each language in the database being parsed using dependency grammar annotations. Resulting units were to be aligned between languages. Translation would be carried out by isolating possible units in the source text, retrieving these units and their 'translation' in the database and combining the retrieved translation units. The work is still somewhat untested, though the work carried out by Tsujii *et al.* (1991) on machine translation by example is similar in spirit to this work and also requires parallel aligned corpora for its operation.

The important point to note here is that again a hybrid model seems to be developing. EBMT does not reject the intuition of the linguist. Similarly it does not eschew the wealth of information available to us from these fabulous modern Rosetta stones, parallel aligned corpora. EBMT seeks a middle path between the cognitively plausible approach of the rule-based systems and the abstraction of statistical MT. It seeks to avoid the high cost of knowledge-base building and the disappointingly poor results of traditional rule-based systems by using the corpus as a means of both supporting the translation process by direct example-based translation, where possible, and as a source of relevant linguistic information, where direct example-based translation is not possible.

On the other hand it also seeks to avoid the abstract non-linguistic processing of pure statistical MT, in favour of a more cognitively plausible approach to the problem. The important point is that, without parallel aligned corpora, this approach to MT would simply not be possible. Corpora have once again provided the raw resources for a new and promising approach to a difficult problem in language engineering.

To return to the general discussion of MT, it is obvious that corpora have had quite an impact in the field of MT. Even where traditional rule-based MT is being undertaken, it may be the case that the system designers may exploit corpora to clear the knowledge acquisition bottleneck for their system in certain key areas (e.g., the lexicon). But where EBMT or statistical MT are taking place the corpus is of paramount importance and is, indeed, a necessity. EBMT and statistical MT have, like so many other areas of language engineering, been enabled by the creation of certain kinds of corpora. Without parallel aligned corpora, there would be no EBMT and there would be no statistical MT.

5.8. CONCLUSION

In general then, what have corpora got to offer to language engineering? From what has been reviewed in this chapter it is apparent that corpora have a great deal to offer language engineering . Apart from being excellent sources of quantitative data, corpora are also excellent sources of linguistic knowledge, in such systems as EBMT.

Indeed, corpora are essential to some applications in language engineering, such as so-called 'probabilistic' part-of-speech annotation systems. Their use in such systems is, however, revealing of the general trend of the use of corpora in language engineering . Although in general these systems are highly effective in terms of reliability, they are also generally of the 'disambiguation' variety. As noted in this chapter, 'radical' applications of corpus data in language engineering are relatively rare. Some variety of disambiguation seems to be the most commonplace use of corpora in language engineering , with few exceptions.

The impact of corpora on language engineering is increasingly profound. At the time of writing, there seems no reason to suppose that this trend will not be amplified on an ongoing basis. The use of corpora in language engineering , especially within hybrid systems where corpus data are used in some process of disambiguation, is burgeoning.

5.9. STUDY QUESTIONS

1. Using the lexicon, suffix list and tagset given below, try to estimate what analyses a part-of-speech analysis system would assign to the sentence *The dog ate the bone*. Disregard case information (*A* as opposed to *a*) when searching through the lexicon, which is all lower case. Then repeat the exercise with *The fat old cat is getting lazy.*

SIMPLIFIED PART-OF-SPEECH TAGSET

A	Article (*the*, *a*, *an* etc.) or Possessive Pronoun (*my*, *your* etc)
C	Conjunction (*and*, *or*, *but*, *if*, *because*)
D	Determiner (*any*, *some*, *this*, *that*, *which*, *whose*)
E	Existential *there*
F	Formulae or Foreign Words
G	Genitive (*’s*)
I	Preposition (*of*, *on*, *under*)
J	Adjective (*young*, *old*)
M	Number or Fraction (*two*, *10’s*, 40–50, *tens*, *fifth*, *quarter*)
N	Noun (*cat*, *cats*, *book*, *north*)
P	Pronoun (*he*, *him*, *everyone*, *none*)
R	Adverb (*else*, *namely*, *very*, *how*, *more*, *alongside*, *where*, *longer*)
T	Infinitive marker (*to*)
U	Interjection (*oh*, *yes*, *um*)
V	Verb (*is*, *was*, *had*, *will*, *throw*)
X	Negator (e.g., *not*)
Z	Letter or letters of the alphabet (*a*, *b*s, *c* or *a*s)
O	Other (e.g., punctuation)

LEXICON

ARTICLES (A)

the a an

PREPOSITIONS AND CONJUNCTIONS (I OR C)

abaft aboard about above across afore after against albeit along alongside although amid amidst among amongst and anent anti around as aslant astraddle astride at atop bar because before behind below beneath beside besides between betwixt beyond but by circa contra despite down during ere ergo except for forasmuch from how howbeit however if in inasmuch inside insomuch into less lest like minus modulo near neath nor notwithstanding o’er of off on once onto opposite or out outside over pace past pending per plus pro qua re round sans save since so than that therefore though through throughout till to toward towards under underneath unless unlike until unto up upon versus via vis-à-vis when whence whenever whensoever where whereas whereat whereby wherefrom wherein whereinto whereof whereon wheresoever wherethrough whereto whereunto whereupon wherever wherewith wherewithal whether while whilst whither whithersoever why with within without worth yet

DETERMINERS (D)

a all an another any both double each either enough every few half last least less little many more most much neither next no none other several some the twice here that there these this those

NUMBER OR FRACTION (M)

zero one two three four five six seven eight nine ten eleven twelve thirteen fourteen fifteen sixteen seventeen eighteen nineteen twenty thirty forty fifty sixty seventy eighty ninety hundred thousand million billion trillion quadrillion quintillion sextillion septillion octillion nonillion decillion zillion

ORDINALS (M)

(irregular only) first second third fifth eighth ninth twelfth (M)

PRONOUNS (P)

anybody anyone anything ’em everybody everyone everything he he’d he’ll he’s her hers herself him himself his hisself i i’d i’ll i’m i’ve it it’s its itself me mine my myself nobody nothing oneself ours ourselves she she’d she’ll she’s somebody someone something their theirs them themselves they they’d they’ll they’re they’ve ’tis ’twas us we we’d we’ll we’re we’ve what what’re whatever whatsoever where’d where’re which whichever whichsoever who who’d who’ll who’s who’ve whoever whom whomever whomso whomsoever whose whoso whosoever yer you you’d you’ll you’re you’ve your yours yourself yourselves

ADVERBIALS (R)

anyhow anymore anyplace anyway anyways anywhere e’en e’er else elsewhere erstwhile even ever evermore everyplace everyway everywhere hence henceforth henceforward here hereabout hereabouts hereafter hereby herein hereinabove hereinafter hereinbelow hereof hereon hereto heretofore hereunder hereunto hereupon herewith hither hitherto ne’er never nohow nonetheless noplace not now nowadays noway noways nowhere nowise someday somehow someplace sometime sometimes somewhat somewhere then thence thenceforth thenceforward thenceforwards there thereabout thereabouts thereafter thereat thereby therefor therefore therefrom therein thereinafter thereinto thereof thereon thereto theretofore thereunder thereunto thereupon therewith therewithal thither thitherto thitherward thitherwards thrice thus thusly too

AUXILIARIES (V)

ain’t aint am are aren’t be been being did didn’t do does doesn’t don’t had hadn’t has hasn’t have haven’t having is isn’t was wasn’t were weren’t can can’t cannot could couldn’t may mayn’t might mightn’t must mustn’t needn’t ought oughtn’t shall shan’t should shouldn’t used usedn’t will won’t would wouldn’t

SUFFIX-MATCHING RULES FOR A SIMPLE PROBABILISTIC TAGGER

Suffix	**Rules**
-s	Remove -s, then search lexicon
-ed	V J@
-ing(s)	-ingV N J@ ; -ings N
-er(s)	-er N V J ; -ers N V
-ly	R J@
-(a)tion	N
-ment	N
-est	J
-(e)n	J N
-ant	J N
-ive	No rule possible
-(e)th	N M
-ful	J N%
-less	J
-ness	N
-al	J N%
-ance	N
-y	J N
-ity	N
-able	J

DEFAULT

If nothing else works, the word in question may be VERB, NOUN or ADJECTIVE (V, N, J).

2. How would you disambiguate the analysis of the sentence from question (1), using your linguistic knowledge?
3. Using the table and equation below, show how we could disambiguate the part-of-speech analysis of the sentence given in question (1).

SIMPLIFIED TRANSITION MATRIX.

This matrix stores a transition probability, going from one part-of-speech (entries along the side) to another part-of-speech (entries along the top).

	A	**C**	**D**	**E**	**F**	**G**	**I**	**J**	**M**	**N**	**P**	**R**	**T**	**U**	**V**	**X**	**Z**	**O**
A	0	0	1	0	0	0	0	29	4	64	0	2	0	0	0	0	0	1
C	14	2	5	2	0	0	9	7	1	19	14	9	1	0	13	1	0	3
D	5	1	3	0	0	0	8	6	2	42	6	3	0	0	16	0	0	7
E	2	0	1	0	0	0	0	0	0	0	0	2	0	0	93	0	0	0
F	0	2	0	0	49	0	2	0	1	7	0	1	0	0	6	0	0	30
G	1	1	1	0	1	0	0	7	1	84	0	1	0	0	0	0	0	2
I	46	0	10	0	0	0	0	6	4	22	4	1	0	0	4	0	0	1
J	0	5	0	0	0	0	5	3	0	72	0	1	4	0	0	0	0	9
M	0	2	1	0	1	0	4	3	2	57	1	2	0	0	2	0	0	26
N	1	6	1	0	0	1	27	0	1	8	1	3	1	0	17	0	0	31
P	1	2	1	0	0	0	4	1	0	1	1	7	2	0	71	0	0	9
R	9	4	3	0	0	0	14	11	1	3	6	11	1	0	21	0	0	16
T	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0
U	1	1	1	0	0	0	1	0	0	1	1	2	0	3	0	0	0	89
V	17	2	3	0	0	0	11	5	1	4	5	16	5	0	18	3	0	10
X	7	1	2	0	0	0	5	4	0	1	6	8	4	0	56	0	0	6
Z	0	4	1	0	2	0	19	2	0	4	0	1	0	0	10	0	0	55
O	9	10	4	0	1	0	5	4	1	10	13	6	0	1	7	0	0	29

EQUATION

For each of the possible tag sequences spanning an ambiguity, a value is generated by calculating the product of the values for successive tag transitions taken from the transition. Imagine we were disambiguating the sequences N followed by a word which is either N or V, followed by a word which is either N or V:

$$\begin{aligned} N-N-V &= 8+17 &&= 25 \\ N-N-N &= 8+8 &&= 16 \\ N-V-V &= 17+18 &&= 35 \\ N-V-N &= 17+4 &&= 21 \end{aligned}$$

The probability of a sequence of tags is then determined by dividing the value obtained for the sequence by the number of sequences, e.g., for the this example the most likely answer is the third:

$$= \frac{35}{2} = 17.5\%.$$

If a tag has a rarity marker (@ for 10 per cent rare, % for 100 per cent rare, then divide the percentage from the matrix by either 10 (@) or 100 (%) before inserting it into the equation.

4. Use the resources provided to analyse sentences taken at random from this book. How easy is it to analyse these sentences using the materials provided? Do any shortcomings of the materials provided become apparent?

5.10. FURTHER READING

A great deal of the work described in this chapter has been published in conferences and journals. As such, picking out easily available key readings is difficult. One text which recommends itself is Black, Garside and Leech (1993), as it covers in detail the shortcomings of various parsing systems, while covering probabilistic part-of-speech tagging, issues in corpus construction and one variety of probabilistic parsing. Another useful book is Zernik (1991), as this contains a variety of papers covering corpus-based approaches to lexicography. Finally, Karlsson *et al.* (1995) is worth reading as it reviews probabilistic part-of-speech taggers, as well as introducing constraint grammars, which have not been covered here.

NOTES

1. Both quotes are from Robert Garigliano, see http://www.dur.ac.uk/~dcs0www3/lnle/editorial.html.
2. Disambiguation: choosing the preferred analysis from a variety of possible analyses. This may occur at many levels, from deciding the meaning, in context, of a polysemous word, through to choosing one possible translation from many.
3. See, for example, http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/.
4. A note of caution should, however, be sounded here. There is evidence that there is an optimal size for such a training corpus. Should the program be exposed to too much training data of this sort, the likelihood seems to be that the model which actually deteriorate rather than improve. See Merialdo (1994) and McEnery *et al.* (1997).
5. It may be that in the literature you will see sentences described as S-units. This is a more neutral term, reflecting the difficulty of unambiguously defining what is and what is not a sentence.
6. For readers interested in seeing a detailed example of this see Markantonatou and Sadler (1994).
7. Needless to say, the probabilistic element is used for ambiguity resolution.
8. Needless to say, these were parallel aligned corpora. It also goes without saying that as a rule of thumb, the larger they are the better. Though recent research has shown that one can overtrain a statistical model, and actually force the performance of a statistical model to degrade by exposing it to a larger sample (Merialdo 1994).
9. Though sentence alignment is much more effective than word alignment.
10. There is another heuristic component to this system, which estimates how often one sentence in language A translates into one sentence in language B, or two in language B and so on and so forth.