1

Introducing Corpus Annotation

GEOFFREY LEECH

1.1 What is a Corpus and What is Corpus Annotation?

Traditionally, linguists have used the term **corpus** to designate a body of naturally-occurring (**authentic**) language data which can be used as a basis for linguistic research. This body of data may consist of written texts, spoken discourses, or samples of spoken and/or written language. Often it is designed to represent a particular language or language variety. In the past thirty-five years, the term **corpus** has been increasingly applied to a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering (see Leech 1991, Leech and Fligelstone 1992, Church and Mercer 1993, McEnery and Wilson 1996). As the power and capacity of computers have increased, corpora have increased dramatically in size, variety and ease of access. At the same time, an increasing range of software has been developed to process corpora and access the information they contain. A computer corpus is fast becoming a universal resource for language research on a scale unimaginable thirty-six years ago.

The mention of a period of thirty-six years is not fortuitous. The year 1961, which more famously saw the first manned space flight, is the date to which corpus linguists can look back as the date when the enterprise now known as **corpus linguistics** (or more precisely **computer corpus linguistics**) came into being. This was the date when work began on the first electronic corpus, later to be known as the Brown Corpus[1] (after Brown University, Providence, RI, where the corpus was compiled). The corpus consisted of just over one million words, comprising 500 text samples of about 2,000 words each. The samples were all taken from publications in the year 1961, and the corpus was complete and ready for distribution on magnetic tape in 1964. As an indication of how the size of corpora has increased since 1964, the one-million-word Brown Corpus seems small today beside the corpus products of the 1990s, including the 100-million-word British National Corpus (BNC),[2] completed in 1994 and containing 10 million words of transcribed speech, and the even larger Bank of English, which runs to more than 300 million words.[3]

However, the value of a corpus as a research tool cannot be measured in terms of brute size. The **diversity** of the corpus, in terms of the variety of registers or text types it represents, can be an equally important (or even more important) criterion. So, too, can the care with which it has been compiled, for example, with respect to the faithful encoding of orthographic features of the text. A fourth factor, the degree to which 'added value' is brought to a corpus by **annotation**, is the subject of this book. Corpus annotation is widely accepted as a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development. Further, as this book will aim to demonstrate fully, corpus annotation has become an important and fascinating area of research in its own right.

But what *is* corpus annotation? It can be defined as the practice of adding **interpretative**, **linguistic** information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with, or interspersed with the electronic **representation** of the language material itself. A typical and familiar case of corpus annotation is **grammatical tagging** (also called word-class tagging, part-of-speech tagging or POS tagging). In this case, a label or **tag** is associated with a word (e.g. by some kind of attachment symbol such as the underline character or the slash character), to indicate its grammatical class: for example, in *taken*_VVN, the grammatical tag VVN shows that *taken* is a past participle. The definition of annotation above, and in particular the use there of the terms 'interpretative' and 'linguistic', requires some further discussion.

First, by calling annotation 'interpretative', we signal that annotation is, at least in some degree, the product of the human mind's understanding of the text. There is no purely objective, mechanistic way of deciding what label or labels should be applied to a given linguistic phenomenon.[4] Disagreement is unlikely to occur if we label *taken* as a past participle – which is conventionally the grammatical class it belongs to in English. But there are many other words which would be more contentious: for example: *future* in *his future bride*. Is it a noun or an adjective? Or, to take up a question of how much detail (**delicacy** or **granularity** are the terms often used for 'detail') should be encoded through annotation, if *future* is an adjective, should it be labelled as an adjective of a particular subclass – say, the class of adjectives which must occur in a pre-nominal position? (We can say *his future bride*, but not **His bride will be future*.) Decisions about these and many other matters have to be taken when we set out to annotate a corpus (see below).

Box 1.1 Example of grammatical tagging, using the C5 tagset of the BNC

High_AJ0 winds_NN2 and_CJC heavy_AJ0 seas_NN2 have_VHB been_VBN
causing_VVG further_AJ0 problems_NN2 in_PRP the_AT0 southern_AJ0
part_NN1 of_PRF Britain_NP0 ,_PUN leaving_VVG homes_NN2
flooded_VVN ,_PUN and_CJC roads_NN2 blocked_VVN ._PUN

Second, we assume a distinction between the 'annotation' and 'representation' of a text – a distinction which may be easy or less easy to apply. For a written text, generally these two kinds of information are relatively easy to separate. The purely orthographic record of a text is a sequence of written characters from (say) the Roman alphabet, interspersed with spaces and punctuation marks (with occasional use of visual material, numerals and 'non-standard' characters such as mathematical symbols). This record can be represented electronically in the computer by special codes and mark-up,[5] and a one-to-one mapping between these and visual symbols can be maintained: the original orthographic document is simply replaced by an unambiguous representation in the form of an electronic document. It is true that some more or less detailed information may be lost in this process – e.g. font and type-size may no longer be retrievable – but this is felt to be allowable if such information is not judged to be essential to the representation of the text as a linguistic phenomenon. In contrast to this, the *annotation* of a text is *meta*linguistic: instead of telling us what the text itself comprises,[6] it gives information *about* the language of the text.

For a spoken discourse, however, it is not easy to distinguish between representational and interpretative information. In rendering speech in written or electronic form (except where the representation is purely instrumental, as in the case of acoustic wave forms), a transcriber must necessarily *interpret* the discourse in the course of representing it. Most transcriptions, as a matter of convenience, incorporate conventionally-spelt words, using phonetic transcription, if at all, only for exceptional pronunciation. But this merely gives superficial readability to speech events whose real nature – physical, linguistic, or social – may be vastly more complex and elusive. Prosodic labelling of stress and intonation, for example, is to some extent dependent on the judgement and expertise of the transcriber (Knowles 1991), as well as on the system of analysis adopted. There is no doubt that prosodic labelling at one level is a **representation** of part of the data of the speech event being transcribed. However, there is equally no doubt that prosodic labelling is in part an **interpretation** of the event through the auditory perception of this or that listener, even where the perceiver is a highly trained phonetician (Pickering *et al.* 1996).[7] For the purposes of this book, we have decided to give some attention to phonetic and prosodic annotation as types of annotation, while acknowledging their in-between status.[8]

Although the distinction between the **raw corpus** (some prefer 'pure corpus') and the interpretative annotations can be somewhat artificial, it is nevertheless a useful distinction, since we should not see annotations as having the claim to reality and authenticity which belongs to the corpus itself. For a written corpus, the text itself is the data (in the etymological sense **data** are 'givens'), and the annotations are superimposed on it. For a spoken corpus, the recording is what is 'given', and it can also be maintained that a bare verbatim transcript of 'what was said' is itself a kind of 'secondary given', that is, a written record without any addition of less reliable, less clearly-definable, information.[9] Beyond these 'givens' it is difficult to go without implicitly taking up some descriptive or interpretative stance towards the data.

1.2 Why Annotate a Corpus?

Why is it important to be able to annotate a corpus?

1.2.1 Extracting information

Corpora are useful only if we can extract knowledge or information from them. The fact is that to extract information from a corpus, we often have to begin by building information in – that is, by adding annotations. The 'raw corpus' in its orthographic form contains no direct information, for example, about grammar – and this can hinder many of the applications to which a corpus can be put. Consider the word spelt *left*. As a word meaning the opposite of *right*, it can be an adjective ('my *left* hand'), an adverb ('turn *left*') or a noun ('on your *left*'). As the past tense or past participle of *leave*, it is a verb ('I *left* early'). *Left* is therefore a very versatile piece of language – but its various meanings and uses cannot be detected from its orthographic form. This is a disadvantage for one of the most salient uses of a corpus in recent years – its use as a resource for lexicography. But if a corpus is successfully grammatically tagged, each occurrence of *left* will be accompanied by a label indicating its word-class. This is a pre-requisite for anyone using a corpus for making or improving dictionaries. To take another example: the word spelt *lead* in English can be either a noun, pronounced /led/, or a verb, pronounced /li:d/. If we

want to create a machine for converting written language into auditory 'spoken' output – a **speech synthesizer** – it is necessary for the synthesizer to distinguish the noun from the verb, if it is to produce a correct pronunciation. Once again, a grammatically tagged corpus would provide the synthesizer with the information it needs.

1.2.2 Re-usability

It might be argued that to extract information of the types mentioned above, there is no need for an exhaustive annotation of a corpus. It might be sufficient to run a clever little program to recognize that, for example, *left* preceding a noun is an adjective, or that *left* following a verb is an adverb. Such little programs could run 'on the fly' extracting instances of the target word without undertaking any annotation. However, such an argument has two weaknesses. First, it is evident from the example of *left* that, in order to identify the word-class of the target word, we would also have to presuppose knowledge of the word-class of neighbouring words. In other words, the identification of word-classes (or any other linguistic phenomena) cannot be treated as an isolated problem. Second, the point about grammatical tagging and other levels of annotation is: once the annotation has been added to the corpus, the resulting annotated corpus is a more valuable resource than the original corpus, and can now be handed on to other users. This argument of 're-usability' is a powerful one, since corpus annotation tends to be an expensive and time consuming business. We do not want to waste resources by 're-inventing the wheel' time and time again – i.e. by re-analysing or re-annotating the same corpus material. An annotated corpus, like any corpus, is valuable because it is a re-usable resource.

1.2.3 Multi-functionality

Taking the point about re-usability one step further, we may note that annotation often has many different purposes or applications: it is multi-functional. We have already noted the application of grammatical tagging to the two different applications of lexicography and speech synthesis. Other language engineering applications – such as machine-aided translation or information retrieval – could also be mentioned. But the general point to make is that annotation gives 'added value' to a corpus in the general sense: it adds overt linguistic information, which can then be used for a multitude of purposes. Thus grammatical tagging is often considered a kind of 'base camp' annotation which can be the first step towards more difficult levels of annotation such as those of syntax and semantics. The reusability of annotated corpora is enhanced by the fact that there are many different purposes for which others may wish to make use of the annotations: purposes which the original annotators of the corpus may not even have thought of.

1.3 Some Standards for Corpus Annotation

Our acceptance of annotations as useful and informative must depend to a considerable extent on our evaluation of the 'experts' who added them to the corpus, and of the usefulness of the annotative scheme they have adopted. In the short history of corpus annotation, it has been by no means unusual for the builders of a corpus to add to it annotations which others have found difficult or impossible to use. To avoid this situation, we suggest that a number of practical guidelines, or standards of good practice, should apply to any project for annotating corpus texts:

1. It should always be possible, and easy, to dispense with the annotations, and to revert to the raw corpus. The raw corpus should be **recoverable**.
2. The annotations should, correspondingly, be **extricable** from the corpus, to be stored independently if there is a need.
3. The user of the corpus[10] should have (easy) access to **documentation**, which will include information on
   (a) The **annotation scheme** – that is, a document describing and explaining the scheme of analysis employed for the annotations.[11]
   (b) **How, where** and **by whom**, the annotations were applied.
   (c) Further, since annotations (given the typical size of annotated corpora) quite often contain erroneous, inconsistent or ambiguous elements, there should be some account of the **quality** of annotation: e.g. to what extent has the corpus been checked, what is its accuracy rate (e.g. the percentage of annotations which are judged correct), and to what extent is the application of annotations consistent (see Chapter 17).

On a more philosophical level, the following additional maxims apply generally both to the compilers and users of annotated corpora:

4. For reasons already given, there can be no claim that the annotation scheme represents 'God's truth'. Rather, the annotation scheme is made available to a research community on a *caveat emptor* principle. It does not come with any 'gold standard' guarantee, but is offered as a matter of practical usefulness only, on the assumption that many users will find it valuable to use a corpus with annotations already built in, rather than to have to devise and apply their own annotations and annotation schemes from scratch (a task which could take years to accomplish).
5. Therefore, to avoid misunderstandings and misapplications, it is good idea for annotation schemes to be based as far as possible on **consensual** or theory-neutral analyses of the data. Perhaps the best analogy here is to the kind of structural or classificatory information given in printed dictionaries. A dictionary gives information about the grammatical classification of words, for example, but tends to take these as given by general descriptive traditions, rather than as coming from some theoretical model that has to be justified. While annotators are bound to face some theoretically sensitive decisions, their goal[12] should be to adopt annotations which are as widely accepted and understood as can be managed. (Perhaps it should be added that the existence and content of 'consensual categories' is not itself a matter on which it is easy to gain a consensus!)
6. No one annotation scheme should claim authority as an absolute **standard**. Annotation schemes tend to vary for good practical reasons. For example, the size of the corpus to be annotated may militate against too much detail. The purpose for which the annotations are primarily intended may give priority to certain kinds of information (e.g. a corpus which has been grammatically tagged mainly as a preliminary to parsing may need careful discriminations to be made between different kinds of subordinating or coordinating conjunction). The kind of corpus data (e.g. spoken vs. written) or the identity of the language (e.g. Chinese vs. Greek) may also encourage differences in the annotations to be applied.

Yet, in spite of (6) above, there is much to said in favour of some kind of standardization of corpus annotation practices, and it is likely that convergence towards some degree of uniformity of practice will take place in the next few years – indeed this convergence has already begun. One reason for standardization is inertia: if you are familiar with some annotation scheme that you have found useful (say, the Penn tagset for grammatical tagging, developed at the University of Pennsylvania – Santorini 1990), it makes sense to stick to that one in developing your own annotated corpus. Another reason is the already-emphasized principle of re-usability. If different researchers need to interchange data and resources (such as annotated corpora), this is more easily achieved if the same standards or guidelines have been applied in different centres. The need for some kind of standardization of annotation practices is particularly evident when we come to the mutual exchange of corpus software utilities (see Chapter 13). Authorities who fund research may also find it desirable to exert influence in the direction of standardization: this has been recently noticed in the policy of the European Union in setting up the EAGLES initiative (see Chapter 16). But the need is to encourage convergent practice without imposing a straitjacket of uniformity which would inhibit flexibility and productive innovation.

1.4 A Glance at the History of Corpus Annotation

1.4.1 Beginnings of grammatical word tagging

To our knowledge, the first computer corpus annotation project to be undertaken was the word-class tagging of the Brown Corpus. Under the supervision of the founders of computer corpus linguistics, Francis and Kučera, this was undertaken by two M.A. students at Brown University, Greene and Rubin (1971), using a tagset of 77 different word-class labels. This was soon after the completion of the Brown Corpus itself. As may be supposed, such a large list of word-class tags would identify not only major parts of speech (noun, verb, preposition, etc.) but also values defining subclasses, such as singular and plural nouns, positive, comparative and superlative adjectives, and so on.

The outcome of this pioneering experiment was that 77 per cent of the words were successfully tagged and disambiguated. (For further discussion see Section 7.1) There still remained the considerable task, undertaken at Brown in the following years, of eliminating all 230,000 of the remaining ambiguities by manual editing of the corpus (see Francis 1980).

The experiment of Greene and Rubin eventually led to an extremely useful product: the word-class tagged Brown Corpus, which has since been used by many thousands of researchers all over the world. But the interest of the TAGGIT method of tagging is that it helps to identify, even at this pioneering stage, a number of general characteristics of corpus annotation. One distinction often made is between automatic and manual annotation of a corpus. Greene and Rubin found it necessary to adopt an automatic tagging method, but the completion of their task was a tedious and time-consuming manual editing of the whole corpus. This division of labour between automatic and manual methods is a recurring theme of corpus annotation, with a number of variations. Beyond a given corpus size (depending on the speed and complexity of annotation), purely manual methods are impracticable. At the other end of the scale, purely automatic annotation can only be tolerated if the result of the annotation is good enough to use as it is: i.e. the error rate or ambiguity rate should be sufficiently low – no more than *n* per cent, where *n* is a small number, dependent on the application – to make the annotated corpus practically useful.

A second major tagging project, in 1979–82, was the tagging of the British counterpart of the Brown Corpus – the LOB Corpus[13] (Marshall 1983; Garside *et al.* 1987: Chapters 3–5). This time the tagging software employed probabilistic methods. Those tagging the LOB Corpus were fortunate enough to be able to use the tagged Brown Corpus as input, especially as a source of tag transitional frequency data. The success rate of automatic tagging leaped from 77 per cent to 96.7 per cent. However, a consequence of the probabilistic method was that the tagger (CLAWS1) inserted the most likely tag in every position, so that wherever it failed, it made errors. That is, 3.3 per cent of the tags were erroneous, and had to be corrected (not merely disambiguated) by hand.

After CLAWS1, a number of word-class taggers were devised, many of them using probabilistic methods (e.g. the taggers of Church (1988) and DeRose (1991)). A number of themes which recur in corpus annotation made their appearance in the decade following the LOB tagging project: the choice between probabilistic and non-probabilistic methods is still a bone of contention. Also, as the LOB tagging project shows, a probabilistic model requires a **training corpus**, a corpus preferably already annotated which supplies initial estimates on the basis of which the probabilistic annotation software is trained. In the case of CLAWS1, this was generously supplied by Kučera and Francis, the authors of the previously tagged Brown Corpus. Another interesting observation is that both TAGGIT and CLAWS, in spite of their different methods of tagging, used a very limited context (one or two words to the left or to the right) to determine the correct tag for a word. This, again, is a recurring issue of corpus annotation software: how far can we get by using extremely local information as a basis for automatic annotation? A final thing to note is that both TAGGIT and its successor CLAWS1 operated on the English language only. For a long time, and indeed up to about 1988, very little annotation of corpora for other languages took place, largely, no doubt, because such corpora did not exist.[14] Since 1990, however, the annotation of corpora has extended to many other languages (e.g. Chinese, Japanese, French, German, Polish, Spanish), and there has even been a move toward the development of language-independent corpus annotation software (especially Cutting *et al.*'s 1992 Xerox Parc tagger – see Chapter 10, especially Section 10.2).

A boom in grammatical tagging began in about 1987, and since that time many taggers have been developed for different languages. Now, however, it is time to backtrack to the mid-seventies to trace the development of other levels of annotation.

1.4.2 *Beginnings of prosodic annotation*

Since written corpora are easier to collect and compile than corpora of spoken discourse, it was not until the mid-1970s that a first major attempt was made to establish a computer corpus of spoken language. This was the London-Lund Corpus (LLC), which was in fact a computer-readable version of spoken materials from the Survey of English Usage corpus (eventually 500,000 words), which had been compiled in paper form at University College London from 1960.[15] The name 'London-Lund' derives from the fact that the computerization was undertaken at Lund, in Sweden (see Svartvik 1990). The LLC was also the first electronic corpus to have prosodic annotation/transcription built into it. The stress, intonation, pauses and other prosodic features had been transcribed in great detail over the preceding 15 years or so in London (see Peppé 1995). Another landmark worth mentioning was the completion in 1986 of the Lancaster/IBM Spoken English Corpus (SEC), which, although much smaller than the LLC, combined different levels of annotation within the same corpus: the same spoken texts were provided with grammatical tagging, syntactic annotation and prosodic annotation, as well as with co-existing orthographic and digitally-recorded versions.[16]

1.4.3 *Beginnings of syntactic annotation*

The mention of the syntactically annotated version of the SEC brings us to another part of the annotation story: the development of corpora with syntactic annotation. In the early days of electronic corpora, a pioneering effort by Ellegård (1978) and his industrious students at Göteborg (Sweden) produced a hand-parsed section of the Brown Corpus. The 'Gothenburg Corpus', as it has been called, consisted of samples amounting to 128,000 words. In the early 1980s, a team at Nijmegen began the TOSCA system for parsing corpus sentences (see van Halteren and Oostdijk 1993), and the team at Lancaster who had tagged the LOB Corpus attempted the parsing of the same corpus by probabilistic methods (Garside and Leech 1985, Garside *et al.* 1987), although hardware and software limitations prevented the completion of the task.[17] In the later 1980s and early 1990s the building of **treebanks** (i.e. parsed corpora – see Chapter 3) took off as a major activity: it was becoming recognized that syntactically annotated corpora were an important resource for the development of NLP software, for example in the development of robust wide-coverage parsers for such applications as speech recognition and machine-aided translation. The Lancaster/IBM treebank (compiled in 1987–91) comprised about 3 million words (Leech and Garside 1991), and the Penn Treebank initiative (Marcus *et al.* 1993)[18] brought the fruits of this new technology to a wider public of users.

The convenient term 'treebank', commonly used for syntactically annotated corpora, brings to notice the fact that the phrase-structure (PS) tree remains the favoured basic model for corpus parsing. Being a more complex and resource-demanding task than grammatical word-tagging, corpus parsing lags behind grammatical tagging in all respects: it began later, it has been less successful, and has been liable to greater inaccuracy, ambiguity, and incompleteness. Early attempts at parsing have had to make do with simplified constituent-structure models, and hence the term 'skeleton parsing' or 'skeletal parsing' was used to characterize the initial Lancaster/IBM and Penn treebanks (Leech and Garside 1991; Marcus *et al.* 1993).

Corpus parsing is still an evolving technology, but it is evolving at a rapid rate. The current state of the art will be further discussed in Chapter 3, but it is worth noting here that, whereas the first treebank (the Gothenburg Corpus) was entirely annotated by hand, we are now reaching a stage where automatic parsing (without extensive post-editing) is becoming practicable. This trend is perhaps best illustrated by the Constraint Grammar parser of the Helsinki group (Karlsson *et al.* 1995), which, although its output is a partial rather than complete parse, does run relatively satisfactorily over large corpora, and has indeed been used to annotate the Bank of English corpus of more than 300 million words. (The Constraint Grammar formalism is also notable for incorporating a dependency grammar framework, in contrast to the PS models employed for most other treebanks.) Towards the other end of the spectrum of size, but equally significant in its way, is the SUSANNE Corpus which is a manual reworking, in considerable detail, of the Gothenburg Corpus, each decision being justified by a detailed parsing scheme published in book form (Sampson 1995). As with tagging, syntactic annotation has a methodological continuum running from 'entirely automatic' to 'entirely manual'. Somewhere on this continuum is the potentially interactive method employed with increasing success by the Nijmegen group (Aarts *et al.* 1993, van Halteren and Oostdijk 1993), where automatic parsing takes place in an environment allowing or requiring intervention to complete the task of satisfactory parsing.

1.4.4 *Other levels*

Although most of the effort in corpus annotation so far has gone into work at the word-class and syntactic levels, other levels of annotation are now beginning to take off: for example, semantic annotation and discoursal annotation. Section 1.5 looks at the different levels of annotation which already exist, summarizing the current state of progress. In Chapters 2–6, these will be explored in greater depth.

1.5 What Levels of Annotation Exist or Can Exist?

Up to now, different levels of annotation have been applied rather patchily, as the list in Box 1.2 (working from the least abstract to the most abstract levels of analysis) indicates. The right-hand column indicates the relevant chapter or section of this book.

As every one of these types of annotation will be discussed and explained in later chapters, the brief illustrations in Box 1.3 are all that are needed at this stage.

Box 1.2 Levels of Corpus Annotation

Linguistic level	Annotations carried out so far	Chapter of this book
Orthographic	This is generally considered part of 'mark up'	(but see §1.5.1)
Phonetic/ phonemic	Widespread in speech science – but typically collected in laboratory situations	(see n.8 this chapter)
Prosodic	Two or three prosodically-annotated corpora are available for widespread use	§6.1
Part of speech (i.e. grammatical tagging)	The most widespread type of corpus annotation, which has been applied to many languages	(Chap. 2)
Syntactic, i.e. (partial) parsing	This is the second most widespread type of corpus annotation, and is rapidly developing	(Chap. 3)
Semantic	Some exists, and more is developing	(Chap. 4)
Discoursal	Little exists – but some is developing	(Chap. 5)
Pragmatic/ Stylistic	(As for discoursal annotation)	(§§6.2–3)

Box 1.3 Brief illustrations of levels of annotation

Example 1a Prosodic annotation, London-Lund Corpus

```
well ^very nice of you to ((come and)) _spare the
!t\/ime and#
^come and !t\alk# -
^tell me a'bout the – !pr\oblems#
and ^incidentally# .
^I [@: ] ^do ^do t\ell me#
^anything you 'want about the :college in ''!g\eneral
```

Example 1b Grammatical tagging from the Penn Treebank, using the Penn Tagset

```
Origin/NN of/IN state/NN automobile/NN practices/NNS ./.
The/DT practice/NN of/IN state-owned/JJ vehicles/NNS for/IN use/NN
of/IN employees/NNS on/IN business/NN dates/VVZ back/RP over/IN
forty/CD years/NNS ./.
```

Example 1c Skeleton parsing (syntactic annotation) from the Spoken English Corpus

```
[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N
who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF [N
his_APP$ pool_NN1 [P on_II [N Clacton_NP1 Pier_NNL1
N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RR [P at_II [N
his_APP$ new_JJ home_NN1 [P in_II [N Windsor_NP1 [ safari_NN1
park_NNL1 ]N]P]N]P]V] ._. S]
```

Example 1d A type of semantic word-tagging

```
There_Z5 's_Z5 been_A3+ more_N5++ violence_E3- in_Z5 the_Z5
Basque_Z2 country_M7 in_Z5 northern_M6 Spain_Z2 :_PUNC one_N1
policeman_G2.1/S2m has_Z5 been_Z5 killed_L1- ,_PUNC and_Z5
two_N1 have_Z5 been_Z5 injured_B2- in_Z5 a_Z5 grenade_G3 and_Z5
machine-gun_G3 attack_G3 on_Z5 their_Z8 patrol-car_M3/G2.1 ._PUNC
```

Example 1e Discoursal Annotation (anaphoric)

```
(0) The state Supreme Court has refused to release {1[2 Rahway State
Prison 2] inmate 1}} (1 James Scott 1) on bail .
(1 The fighter 1) is serving 30-40 years for a 1975 armed robbery
conviction . (1 Scott 1) had asked for freedom while <1 he waits for an
appeal decision. Meanwhile , [3 <1 his promoter 3] , {{3 Murad
Muhammed 3} , said Wednesday <3 he netted only $15,250 for (4 [1
Scott 1] 's nationally televised light heavyweight fight against {5 ranking
contender 5}} (5 Yaqui Lopez 5) last Saturday 4) .
```

1.5.1 Orthographic annotation

Orthographic annotation might seem to be a contradiction in terms – since, as we have seen, orthography represents the text, while annotation interprets the text linguistically. However, up to a point orthographic information can be interpretive, in distinguishing the linguistic functions of various visual devices on paper. Consider different graphological signals for indicating the beginning and the end of a quotation: single quotes, double quotes and change of typesize accompanied by indentation. These are different in form, but alike in function. Conversely, we can also say that the single mark (') is unitary in form, but ambiguous in function: it can signal a single closing quote, or it can represent an apostrophe. Hence, when we read *friends'* out of context, we have to keep both possibilities in mind. Against this background, the TEI Guidelines (Sperberg-McQueen 1991, Sperberg-McQueen and Burnard 1994)[19] allow us to use a pair of symbols `&bquo;` ('begin quote') and `&equo;` ('end quote') which signal these orthographic functions irrespective of the typographical device used.[20] This TEI mark-up may be regarded as a kind of orthographic annotation. Other ambiguous orthographic devices which might be annotated to resolve ambiguities are:

1. *Initial* capital letters, which may signal the beginning of a sentence, the beginning of a proper noun, etc.
2. *A period* (.), which may signal either the end of a sentence or an abbreviation
3. *Italics*, which may signal a cited expression, an expression italicized for emphasis, etc.

In some cases, these distinctions have been made in the encoding of the orthographic record of a text, and have hence made a useful contribution to corpus processing and annotation at more abstract levels. For example, in the LOB Corpus the symbol `\0` was used to signal a one-word abbreviation (Johansson *et al.* 1978), so that, for example, `\0in.` as an abbreviation for *inch* could be automatically distinguished from the preposition *in* at the end of a sentence. However, in general such orthographic annotation has not been consistently applied to corpora, so that it would be unwise to rely upon it in designing software for corpus annotation.

1.5.2 Additional types of annotation

In addition to the above levels of annotation, there are some other levels which should be mentioned, although they remain largely undeveloped.

First, there are some levels of linguistic structure or function which could be indicated by annotation, although we know of no generally-available corpora annotated at these levels. For example, at the phonological level, corpora could be annotated by syllable boundaries. At the morphological level, corpora could be annotated by their morphological structure, in terms of prefixes, suffixes and stems. Previous experience suggests that, even if no need for such levels of annotation has yet appeared, such a need is quite likely to arise in the future.

Second, there is the level of **lexeme** annotation or **lemma** annotation (these are alternative names for the same concept). When we have talked of grammatical word tagging previously, we have been assuming that, for example, *eat, eats, ate, eaten* and *eating* receive different tags according to their morphosyntactic function as past tense verb, *-ing* form of the verb, etc. However, another approach to grammatical word tagging would give each of these the same tag, and indicate that they all belong to the lemma EAT ('lemma' being more or less equivalent to the headword of a dictionary entry). In English, lemma annotation may be considered somewhat redundant,[21] since English is a language with simple inflectional morphology. But in more highly-inflected languages, such as Russian or Spanish, there is a relatively large number of word-forms per lemma, so that lemma annotation may have a valuable contribution to make to information extraction – for example, for the improvement of dictionaries or computer lexicons of the language.

Third, there is a kind of annotation which does not depend on the simple recognition of different levels of linguistic function, but is more closely geared to applications. Thus, there have recently come into being a number of **learner corpora** of English, representing the language of those learning English as a second or foreign language (e.g. Granger 1993). The function of such corpora is to advance our knowledge of how languages are learned as a second language: for example, to what extent does the English of non-native speakers reflect the influence of their native tongue? For this kind of investigation, it is very useful to annotate the corpus with classes of errors, or features of non-native language behaviour. Such 'error tags' make use of grammatical and lexical classifications, for example, but also take into account the relation between the non-native and corresponding native phenomena.

'Error tagging' of learner corpora is just one example of application-oriented annotations, and there may be many more. This is sufficient to indicate that annotation is an open-ended area of research, which is very much under development. While in the next five chapters we review levels of annotation which already exist, it cannot be doubted that new kinds of annotation will arise in the future.

Notes

1. The corpus was originally more verbosely labelled 'a standard sample of present-day edited American English for use with digital computers' (see Francis and Kučera 1964).
2. As the BNC will be a focus of discussion in a number of chapters, it will be useful here to add some details of its compilation and composition. The BNC is a corpus of 100 million words, containing texts taken from sources such as newspapers, books, magazines, and transcribed conversations, lectures, meetings and interviews (Burnard 1995). The corpus is also annotated, in that individual words have been tagged to show part-of-speech (POS) information. The whole of the BNC has been tagged using the relatively small C5 tagset (see Appendix III) while 2 million words of the corpus, known as the sampler corpus, have been tagged using an expanded version of the tagset, known as the C7 tagset (consisting of 146 POS tags). The corpus was built by a team consisting of Oxford University Press, Longman Group Ltd., Chambers Harrap, The British Library and the Universities of Oxford (Oxford University Computing Services) and Lancaster (UCREL). Further details of the BNC are provided by Burnard (1995); a broad survey is given by Leech (1994).
3. Information on the Bank of English can be accessed on the World-Wide Web at http://titania.cobuild.collins.co.uk/boe_info.html.
4. In fact, this is not quite true. Experiments have been carried out to induce linguistic word classes from a corpus purely automatically, on a distributional basis (see, e.g., Atwell and Elliott 1987), making no use of humanly-devised categories. Such classes sometimes have an uncanny resemblance to categories used in grammatical or semantic tagging (e.g. prepositions, modal auxiliaries, nouns for months). Whether a labelling of corpus words according to these induced categories would be considered a kind of linguistic annotation is a matter of terminological definition.
5. Much of this encoding is purely conventional: the ASCII code is the encoding system generally used for converting the symbols on the terminal or typewriter keyboard to binary electronic form.
6. There is, of course, more than this to the issue of 'what is the purely orthographic record of a text'. Some compilers of corpora have been content with the 'plain ASCII text', without mark-up indicating such linguistically-relevant details as headings and highlighted expressions. Going to the other extreme, others will take the view that any diagrams, photographs, etc., accompanying a written text are as much a part of it as the words themselves. However, these are still issues of what comprises the representation of a text: they do not trespass into the 'metalinguistic' territory of corpus annotation.

   Another kind of information provided in a corpus may be considered distinct from both the text itself and the annotation of the text. This is **header information** (so-called because it tends to be provided in headers, or headings, at the beginning of a text or corpus). This gives information of various kinds about the 'documents' or texts which comprise a corpus, as well as about the corpus in its entirety. Such information may include bibliographical details of a written text, and parallel information about a spoken discourse (regarding identity and background of speakers, the provenance of the transcription, etc.). It may also provide a classification of the 'document' in terms of the text typology used in designing the corpus, hence giving information of an interpretative, linguistic nature – for example, indicating something of the style of language found in the 'document'.

7. For spoken discourse, papers by Ochs (1979) and Cook (1995) deal with issues connected with the non-objectivity of theory. Their papers are provokingly, though aptly, named 'Transcription as theory' and 'Transcribing the untranscribable'.
8. However, we do not attempt to cover in this book the subject of **speech corpora**, by which is generally meant recorded and annotated speech data collected under 'laboratory' conditions, i.e. conditions not comparable to those of authentic spoken discourse. The immense amount of recent work on speech corpora, including annotation, can be seen by consulting the EAGLES World Wide Web site http://coral.lili.uni-bielefeld.de/~gibbon/EAGLES/. A further source of information is the handbook of the EAGLES Spoken Language Standards and Resources Group due for publication in 1997. See also Section 6.1.2 (1).
9. It is notable that for official purposes in society at large, such as the transcription of court proceedings, a verbatim transcript corresponding faithfully to the words spoken, in their right order, is considered to be a faithful record of what was said.
10. Here we are making an assumption that the **user** of an annotated corpus is not the same as the **annotator**. It is possible, of course, that some annotations are done by researchers purely for their own use, with no intention of distributing their annotations to others. However, as far as this book is concerned, the reason why annotation is worth studying in depth is that in natural language processing (NLP) it is increasingly becoming important to re-use the resources compiled or devised by others (see Section 1.2.2). For us, then, the users of a corpus comprise a potentially large group, typically distributed across the world, and engaging in many different kinds of research and development activity.
11. Sampson (1995) and Johansson (1986) are two detailed examples of what an annotation scheme should attempt to do. An annotation scheme should include: (i) a list of the annotative symbols used, (ii) their definitions, and (iii) the rules or guidelines that have been used in their application. Another way in which an annotation scheme can explicate the nature of the annotations is to cross-refer (for instance) to a lexicon or a grammar or a 'reference corpus' which exemplifies the various descriptive decisions made by the annotators.
12. In the interests of **re-usability** (see Section 1.2.2).
13. 'LOB' Corpus is an abbreviation for the Lancaster-Oslo/Bergen Corpus, compiled at the three universities mentioned in its name during 1970–78 (see Johansson *et al.* 1978).

14. Early work was undertaken on the tagging of Swedish at Lund and of Dutch at Nijmegen.
15. The Survey of English Usage project was announced and described in Quirk (1960). The majority of the spoken materials of the LLC have also been published in book form (Svartvik and Quirk 1980).
16. See Knowles (1993). The SEC was later reworked as a CD-ROM where all levels of annotation were combined in a single database, and were cross-registered to the digital soundtrack and the F_0 waveform.
17. The parsing of about 144,000 words of the 1-million-word LOB Corpus was eventually completed, and made available via the Norwegian Computing Centre for the Humanities, under the title of the 'Lancaster Parsed Corpus'.
18. A subset of the Penn Treebank is available to researchers for non-commercial purposes, on payment of a license fee, from the Linguistic Data Consortium (LDC). For further details, see the relevant items on the LDC's World Wide Web site: http://www.ldc.upenn.edu/.
19. 'TEI' stands for the Text Encoding Initiative, an international initiative to set up a flexible standard for the encoding or mark-up of texts for electronic interchange. For most purposes, we may see the TEI as systemizing the representation of raw text, rather than as being concerned with annotation practices. However, there is a sense in which TEI mark-up is an aspect of annotation practices: it lays down guidelines for the representation of annotations. Just as the raw corpus needs to be represented electronically, so the annotations need to be represented electronically. And it is this aspect of annotation practices (and not, say, the choice of categories) which comes within the purview of the TEI (see further Section 2.4).
20. It needs to be said that SGML, the language which TEI uses, attempts to mark up an original text by **function** rather than **realization**. Thus a word to be emphasized is to be marked as such, rather than as **italic**, and the realization of emphasis (as well as, say, foreign words) by italics would be specified independently.
21. However, lemma annotation has been undertaken by Fligelstone (1995) for a corpus of English newspapers and by Sampson (1995) for the SUSANNE Corpus.