Proceedings Paper:
Nancarrow, O and Atwell, ES (2007) A comparative study of the tagging of adverbs in modern English corpora. In: Proceedings of the CL'2007 Corpus Linguistics Conference. CL'2007 Corpus Linguistics Conference, 24-30 May 2004, University of Birmingham, UK. UCREL, Lancaster University.

A Comparative Study of the Tagging of Adverbs in Modern English Corpora

Owen Nancarrow[1] and Eric Atwell[2]

Introduction

The tagged Brown, tagged LOB, BNC Sampler, and ICE-GB corpora of modern English are valuable resources for the empirical study of English grammar, as they have all been enriched by part of speech tagging.. The Brown Corpus in particular has also been widely taken up by Computational Linguistics and Machine Learning researchers, who use the tagged texts as ML training standard datasets, taking for granted the validity of the tagging. This paper examines closely the tagging of adverbs in these four corpora, and identifies some weaknesses, which show that this assumption about the accuracy of the tagging is not always fully justified. In particular, the Brown Corpus seems to have the largest number of inconsistencies in the tagging of adverbs, so that Machine Learning research would benefit from a switch to another tagged corpus such as LOB with fewer internal inconsistencies. As well as shortcomings in the tagging, some inaccuracies in the Brown Manual have been found and are here discussed.

The tags for adverbs are discussed in relation to contemporary grammatical descriptions of English, with which the tagging schemes clearly have a close relation. Nevertheless, the relationship between pre-existing grammatical categories and tagsets is very different in each corpus, and even within an identical overarching framework, each corpus has its own distinctive taxonomy. The tags have been grouped into related sets, and the tags which are used for classes of words which can justifiably be considered fundamentally similar are further grouped together. Changes in the tagging are noted, and the reasons or lack of reasons for such changes are discussed. Type-token frequency tables are given for all adverb and adverb-related tags examined in this paper. An indexed synoptic table is provided at the end of the paper.

The Brown Corpus is the foundation of all that follows. The reader will see how much the tagging of this corpus has determined the tagging of at least the next two. The final corpus, ICE-GB differs from the previous ones, just as they differ from one another. This corpus, however, has a tagging scheme which contains a number of innovations not found in previous corpora, and these are discussed in this paper. In clarity and consistency of tagging the LOB and Sampler corpora are shown to improve on Brown, as is only to be expected.

Other studies have included comparisons between English corpus tagsets (eg van Halteren 1999, Atwell et al 2000, Jurafsky and Martin 2000), but none to our knowledge has focused on adverbs, or examined differences of sub-categorizations in such detail. The

[1] School of Computing, University of Leeds
e-mail: otnancar@hkucc.hku.hk
[2] School of Computing, University of Leeds
e-mail: eric@comp.leeds.ac.uk

approach in this paper provides a methodology to follow in examining sub-categorizations in other corpus tagsets, and/or other grammatical categories.

1. Related corpora

Any given corpus often belongs to a set of related corpora. For example, an untagged corpus may later be tagged, a smaller corpus may be constructed from a larger one, and new corrected or improved versions may be published. The discussion of adverb tagging in this paper centres on just four corpora, each one belonging to a different set. Before listing them in Table 3, we introduce some basic information about the four sets to which they belong, followed by a brief historical survey.

1.1 The four sets of related corpora relevant to this paper

The four columns of Table 1 identify: (A) the row number, (B) the name of the first published member of each of set, (C) the abbreviation which is usually used for it, if there is one, and (D) its date of publication. A different colour is used for the members of each set: yellow for Brown, orange for LOB, pink for the BNC, and blue for ICE-GB.

A	B	C	D
1	The Brown Corpus	Brown	1964
2	The Lancaster-Oslo-Bergen Corpus	LOB	1978
3	The British National Corpus	BNC	1995
4	The International Corpus of English, British Component	ICE-GB	1998

Table 1: Four corpora of modern English

As new members of the set were added, these simple abbreviations were no longer enough to identify the intended corpus correctly. Table 2 lists the principal published members of each set. Some corpora are simultaneously published in different versions, containing different markup, so that even these abbreviations may need elaboration if the exact corpus version being referred to is to be correctly identified. The seven columns identify (A) the row number, (B) the corpus name, (C) whether the corpus is untagged, tagged, or tagged and parsed, (D) if it is tagged, the tagset name, (E) an abbreviation for easy identification, (F) the approximate size of the corpus in millions of words, and (G) the date of publication.

A	B	C	D	E	F	G
1	Brown	untagged	none	Brown1	1	1964
2	Tagged Brown	tagged	Brown	Brown2	1	1979
3	LOB	untagged	none	LOB1	1	1978
4	Tagged LOB	tagged	LOB	LOB2	1	1986
5	BNC	tagged	C5	BNC1	100	1995
6	BNC World	tagged	C5	BNC2	100	2000
7	BNC XML	tagged	C5	BNC3	100	2006
8	BNC Sampler	tagged	C7	BNCS	2	1999
9	BNC Baby	tagged	C5	BNCB	4	2004
10	ICE-GB	tagged and parsed	ICE	ICE-GB1	1	1998
11	ICE-GB 2nd edition	tagged and parsed	ICE	ICE-GB2	1	2006

Table 2: Four sets of related corpora

1.2 Historical survey

Forty-five years ago, in 1962, the first steps were taken to create a one million word computer corpus of American English, the Brown Corpus, named after Brown University in America where it was created. A few years after it was published in 1964 it was decided to tag the corpus, but the completed and fully-tagged corpus was not ready for publication in its final form until 1979.

Two British linguists, Randolph Quirk and Geoffrey Leech, were involved with the Brown corpus from its very inception, and Geoffrey Leech soon initiated and planned a one million word British English corpus, later called the LOB Corpus from the three places involved in its creation, Lancaster in England, and Oslo and Bergen in Norway. The untagged version was published in 1978, the tagged version in 1986.

Both these corpora were created with the help of automatic tagging programs, then manually corrected with great care and thoroughness. The resulting corpora are therefore believed to be almost free of erroneous tags. However, in the case of the Brown Corpus, this belief is not entirely justified, as we shall see below.

Not too long after the completion of the LOB Corpus, a mighty new project was conceived: not a one million, but a one hundred million word corpus would be created - the British National Corpus, or BNC. Because of its size, all the tagging was done automatically, without further manual correction.

Since the automatic tagging program produced many errors, a two million word subset was selected for the manual correction which was considered impractical for the

parent corpus. This corpus was called the BNC Sampler Corpus. It was also tagged with a larger, more detailed tagset, called C7, rather than with the simpler C5 tagset used for the BNC, The tagging is claimed to be almost error-free. This new corpus was published in 1999.

In 1995, another British linguist, Sidney Greenbaum, who had worked closely with Quirk and Leech, conceived not just a single new corpus, but a whole set of corpora spanning the globe, the International Corpus of English, or ICE, project. The first of these, ICE-GB (GB for Great Britain) was published in 1998. A second edition has just appeared. Although only one million words in size, they were to be not only fully tagged, but also fully parsed. The sentences and their associated tagged and parsed diagrams can be beautifully viewed using a marvellous facility provided with the Corpus termed the ICE Corpus Utility Program or ICECUP.

1.3 The three full corpora and one subcorpus discussed in this paper

Today's paper examines the tagging of adverbs in three corpora and one subcorpus, all about one million words in size. The three complete corpora are Brown2, LOB2, and ICE-GB1, the subcorpus is that half of BNCS - the BNC Sampler - which consists of written rather than spoken English. We shall refer to this subcorpus as BNCS-W. The Brown and LOB corpora, and the Sampler Subcorpus consist entirely of written English, only the ICE-GB Corpus has a mixture of spoken and written English. In fact, the spoken component is 60 percent of the whole, and thus considerably larger than the written component.

Table 3 lists these four corpora. In this and the remaining tables, each corpus is identified not only by a colour, but also by a single letter abbreviation: B, L, S, and I. In the table below, column (A) gives the row number, (B) the usual corpus abbreviations (C) a simpler abbreviation used in the text of this paper, (D) an even simpler one used in tables, and (E) the date of publication.

A	B	C	D	E
1	Brown2	Brown	B	1979
2	LOB2	LOB	L	1986
3	BNCS-W	the Sampler	S	1999
4	ICE-GB1	ICE	I	1998

Table 3: The three corpora and one subcorpus referred to in this paper

2. The grammatical background

The authors of the Brown Corpus explicitly link the tags they use to the classifications of words found in what they call "traditional" grammar. In Section 4 of the on-line Brown Manual they write:

> On the whole the taxonomy is traditional and should be transparent to the grammarian
>
> (Francis and Kucera, 1979: section 4)

The word "traditional" is often used, as it is here, for any descriptive rather than theoretical grammar of English. A typical traditional taxonomy is found in *A practical English Grammar* by Thomson and Martinet , first published in 1960 by OUP, still available today, but now in its fourth edition. In Section 63, seven kinds of adverb are listed under the headings: manner, place, time, frequency, degree, interrogative and relative. In section 90, another set of relevant words is identified: words which can be used either as adverbs or prepositions. Examples of adverbs in each of these categories are given in the quotations below:

> Section 63
>
> There are seven kinds of adverbs:
>
> 1 of manner: e.g. *quickly, bravely, happily, hard, fast, well*
> 2 of place: e.g. *here, there, everywhere, up, down, near, by*
> 3 of time: e.g. *now, soon, yet, still, then, today*
> 4 of frequency: e.g. *twice, often, never, always, occasionally*
> 5 of degree: e.g. *very, fairly, rather, quite, too, hardly*
> 6 interrogative: e.g. *when? where? why?*
> 7 relative: e.g. *when, where, why*
>
> (Thomson and Martinet, 1969: 38)

> Section 90
>
> Some words can be used as either prepositions or adverbs ... The most important words of this type are: *in, on, up, down, off, near, through, along, across, under, round*
>
> (Thomson and Martinet, 1969: 52)

Although there is no tag in any corpus for frequency adverbs, we will use the other headings, together with the adverb-or-preposition category, as a simple, but convenient framework for the first part of this paper.

The traditional taxonomy of books like these was enormously enriched by an exciting new publication in 1972 - *A Grammar of Contemporary English*. The four authors - Randolph Quirk, Geoffrey Leech, Sidney Greenbaum and Jan Svartvik – were all associated in some way with one or more of the four corpora we shall refer to today. This new grammar was itself inspired by working with corpora, and in turn inspired new ideas about how tagsets should be structured. Corpora and grammar development have usually gone hand in hand, and undoubtedly will continue to do so. In 1985, a major enlargement and revision of this already very large grammar appeared - *A Comprehensive Grammar of the English Language*, perhaps the most important "traditional" grammar of English in the second half of the twentieth century.

2.1 Adverb tags

Before we consider the tags themselves, we must say something about which tags we shall include in our discussion. Some tags are clearly labelled as some kind of adverb tag in all four corpora, and these must obviously be included. However, some tags are labelled adverb tags in some corpora but not in others. If a tag is clearly labelled an adverb tag in at least one corpus, then we include this tag, together with the equivalent tags in other corpora, even if the other tags are not labelled as adverb tags. Only one tag is included which is not labelled as an adverb tag in any corpus – that is the tag for certain uses of the word *there*. In table 31 at the end of this paper you can see (almost) all the tags which we shall discuss.

3. Manner adverbs

By far the largest subclass of adverbs, is the subclass called "manner adverbs" in Thomson and Martinet. This subclass is referred to in the corpus manuals simply as the "adverb" or "general adverb" subclass. Table 4 lists the tags, the type totals, the token totals, and the descriptive name used in the relevant manual. This arrangement is followed in the other adverb subclass tables below.

	Tag	Type total	Token total	Descriptive name
B	RB	1,764	36,602	adverb
L	RB	1,688	35,357	adverb
S	RR	1,473	28,207	general adverb
I	ADV(ge)	1,483	32,723	general adverb

Table 4: Manner adverbs

Most of the words in this subclass are derivative *–ly* adverbs, by far the largest group of adverbs. The suffix *–ly* is highly productive, and each of the four corpora has many word types with this suffix which are not found in any of the others.

Before we consider the other subclasses, we shall say a few words about the structure of the tags themselves, and about how the totals were calculated. You might also like to calculate percentages, so some further information allowing you to do this is given at this point.

3.1 The structure of adverb tags

An adverb tag, or indeed any tag, in the first three corpora consists of from two to four capital letters, each letter (or pair of letters) identifying some property of the subclass. The relevant tags can be seen in table 31.

In these three corpora, the adverb tag begins with the letter R. The Brown Corpus was the first to use R as an adverb identifier, and as in so many things, what the Brown corpus did has continued into later corpora. More generally, R as first or second letter of a tag in these corpora always identifies some kind of adverb tag.

In the ICE corpus, a tag has a quite different structure: it consists of a primary class label in capital letters, followed by one or more features in lower case and within parentheses. The features identify subclass properties: For example, the feature *ge*, an abbreviation, identifies the *general* subclass of the primary class of adverbs, abbreviated *ADV*.

No further comments will be made about the adverb tags in subsequent tables, but you can readily find detailed information from the manuals listed at the end of the paper.

3.2 How the totals have been calculated

All the type and token totals in this paper have been calculated after changing the capital letters A to Z to small letters.

By "word type" in this paper is meant a word together with its accompanying tag. In the Brown Corpus cited words are marked by a tag extension *-NC*. This has been removed before calculating the type-token totals. Since any word may be so marked, including it would potentially double the number of tags.

In the LOB Corpus the abbreviation marker *\0* has been removed. In the case of ICE, two kinds of markup included in textual items have been changed: the line-end marker has been deleted, and the line-end hyphen marker has been replaced by a hyphen. In this way, words which would have belonged to different word types had the markup been retained, now belong to the same word type.

All the totals have been calculated directly from data files, using software written by ourselves. This is a first attempt, and although we have tried our best to avoid errors, they always seem to be present. We would be most grateful if readers would inform us of any errors that they may find.

3.3 How to calculate percentages

Before any percentages can be calculated some further totals are needed, and these are given in table 5 below. This table not only gives the type and token totals for each of the four corpora, but also shows the type and token totals for the adverb tags in each corpus.

Punctuation tags have been excluded from the corpus totals below, and in the case of ICE, items with PAUSE tags have not been included either, since there is no comparable tag in the other corpora.

	Corpus type total	Corpus token total	Adverb type total	Adverb token total
B	55,631	1,014,203	2,269	66,482
L	56,043	1,013,729	1,982	76,236
S	56,188	1,002,820	1,824	61,275
I	73,489	992,136	2,565	104,284
Ia			2,331	74,429

Table 5: Type and token totals

For the first three corpora, the adverb tags included in the totals are exactly those included in the synoptic table, table 31, at the end of the paper. The totals are therefore equivalent to the sum of the totals for the tags in this table. The only significant tags omitted from any total are the combined tags of the Brown Corpus discussed in sections 12, 12.1, and 12.2.

In the case of ICE, two sets of adverb figures are given and explained below. In both sets all relevant tags beginning ADV, CONNEC, EXTHERE, and REACT have been included. Tags with the feature *ignore,* discussed in section 15, and the anomalous tags discussed in section 16 have also been included.

The corpus type total is much higher in ICE than the other corpora. A possible contributory factor is probably the fact that noun sequences in ICE have generally been tagged as single words.

In ICE, the adverb token total is very much higher than in other corpora. In this case, the most likely cause is the fact that a number of high frequency words tagged as general connectives or reaction signals (see sections 9.3 *ff.*) are not tagged as adverbs in the other corpora. Excluding the totals for these two tags gives figures much closer to the other corpora. This has been done in the second ICE row, labelled Ia.

The token totals in table 4 and other such tables are from a total word count of about one million. The approximate percentage of such tokens is therefore obtained by placing a decimal point before the fourth digit to the left: for example, the percentage of adverb tags in the Brown Corpus is 3.6602. If there are less than four digits add zeros to make four digits then place the decimal point before it: thus, if there are nine tokens, these constitute approximately 0.0009 percent of the total number of tokens.

4. Adverbs of time and place

Examples of adverbs belonging to this class may be found in the quotations in section 2. Table 6 lists tags for adverbs of time and place.

a Nominal adverbs				
B	RN	3	9	nominal adverb
L	RN	15	4,333	nominal adverb
b Adverbial nouns				
B	NR	26	1,894	adverbial noun
L	NR	60	2,916	singular adverbial noun
c Time adverbs				
S	RT	33	4,635	quasi-nominal adverb of time
d Place adverbs				
S	RL	121	3,106	locative adverb

Table 6: Adverbs of time and place

It can be seen at once that Brown, LOB, and the Sampler have two such adverb subclasses, and that ICE has none at all.

The two Sampler subclasses explicitly recognize classes of time and place, but the Brown and LOB names make no reference at all to either time or place. Nevertheless, each Brown and LOB subclass contains only time and place adverbs, although both types of adverb occur together in each subclass.

If you examine the figures in this table, you will have to conclude that there is no straightforward relationship between subclasses in different corpora, even if they have the same name. Why does the nominal adverb class in Brown have three types and nine tokens, but the subclass with the same name in LOB have fifteen types and 4,333 tokens? Moreover, you might wonder why the combined totals for time and place adverbs in all three corpora are so very different: in Brown there are just twenty-nine, in LOB seventy-six, in the Sampler 156, and in ICE there are none at all.

4.1 Inconsistencies in the Brown Corpus

The first of these questions has a simple, but surprising answer. The other question requires detailed and careful comparisons, which will be undertaken in a thesis being written by one of the authors.

To answer this first question a few preliminaries are necessary. Firstly, it will be agreed that if the same word, with the same meaning, is tagged with more than one tag in identical or very similar contexts then only one tag can be correct. Indeed corpus authors always request users of their corpora to notify them of erroneous tags.

Secondly, a corpus manual should not make incorrect statements about the corpus it describes. Nor should it make contradictory statements.

In relation to the nominal adverb tag RN the Brown corpus fails on both these counts. The following sections show why this is the case.

Table 7 shows the three word types tagged RN, together with their token totals. There are no others.

		token total
1	afar	2
2	here	4
3	then	3

Table 7: Words tagged as nominal adverbs in Brown

But section 4 of the on-line Brown Manual contains the following statement:

> Certain adverbs, mostly temporal or locative, which often function as nominals have been denominated "nominal adverbs" and tagged RN; thus *here, then, indoors*.
>
> (Francis and Kucera, 1979: section 4)

Contrary to what is said, there is no instance of *indoors* tagged RN. And contrary to what is implied by the phrase "Certain adverbs, mostly temporal or locative", the three words in Table 7 are the only words tagged as nominal adverbs.

Moreover, there are numerous instances of *here* and *then* tagged RB in exactly the same environments to those where the tag RN occurs, and hundreds of instances in similar environments. There are only two instances of *afar* in the corpus, but similar words in similar environments are tagged RB. Table 8 gives some typical examples of the inconsistent tagging of these words. Given the large number of tokens of these words, identical or almost identical environments were not hard to find.

		token total
1a	outta here_RN	2
1b	out of here_RB	12
2a	around here_RN	1
2b	around here_RB	10
3a	from here_RN	1
3b	from here_RB	13

Table 8: Inconsistent taggings in Brown

The only reasonable conclusion is that all these words should be tagged RB, and that the tag RN has no place in the corpus.

The fact is that the Brown Corpus Manual more accurately describes the situation in the LOB Corpus. How and why these few words were, as we have argued, wrongly tagged in this way is not known. In any event, to the best of our knowledge, after 1979 no further revisions were made to the tagged Brown Corpus, so that errors existing at that time remained uncorrected.

4.2 Ambiguous tags and differing dictionary labels

The tags nominal adverb and adverbial noun are not like most other tags: their meaning is ambiguous. In any given context, a word tagged with one of these tags is either a noun or an adverb, but cannot, of course, be both at the same time. By inventing and using such tags, the authors are avoiding having to decide which tag is the appropriate one. The Sampler and ICE have no such tags, instead they have resolved the ambiguity, selecting in each case either a noun or an adverb tag for words which have these ambiguous tags in Brown and LOB.

Why did the authors of these corpora create these tags? In the case of the Brown Corpus it was perhaps to avoid having to manually correct errors made by a poor automatic tagging program. In the case of the LOB corpus, it was perhaps the acceptance of a pre-existing Brown tag and, again perhaps, a reluctance to have to make a decision about the grammatical status of such words. Dictionaries did and do often vary in their labelling of these words: the word *now,* for example, is labelled variously as an adverb and noun, as an adverb and pronoun (but not as a noun), and as an adverb (but not as a noun or pronoun).

5. Adverbs of degree

We move on now to a most interesting set of tags, those for degree adverbs. Table 9 shows the relevant tags and totals. The Brown and LOB Corpora each have two identically named subclasses, the Sampler and ICE on the other hand just one subclass each.

a Which may not follow the head				
B	QL	374	8,750	qualifier
L	QL	17	5,375	qualifier
S	RG	39	3,881	degree adverb
I	ADV(inten)	299	12,860	intensifier adverb
b Which may follow the head				
B	QLP	4	263	post-qualifier
L	QLP	2	283	post-qualifier

Table 9: Degree adverbs

The post-qualifier subclasses in Brown and LOB are essentially for just two words, *enough* and *indeed*. In the Sampler these two words, in identical contexts, are tagged as general adverbs, in ICE they are both tagged as intensifier adverbs.

The problem is: why is the tagging different? What are the reasons? It seems that detailed explanations for the choice of particular tags are rarely given. These two words are clearly degree adverbs, and clearly they differ from other degree adverbs in being placed after the word they qualify. Brown and LOB incorporate both properties in their tags, the Sampler ignores both properties, and ICE recognizes only the intensifier property.

The interesting thing is the varying number of word types with the main degree adverb tag: this starts with a large number in Brown, goes right down to a very small number in LOB, stays a small number in the Sampler, and finally in ICE it is back again to a large number. Moreover, the four tag sets fall neatly into two very similar pairs: Brown and ICE, and LOB and the Sampler. After twenty years, and two alternative treatments, the same set of 300 or so adverbs has again been assigned to a single adverb subclass.

5.1 More inconsistencies in the Brown Corpus

In spite of this resemblance between Brown and ICE, there are once again the same problems with inconsistent taggings, and incompatible statements about the Brown Corpus by its creators.

The following statement from section 4 of the on-line Brown Manual tells us that only three adverbs ending in *-ly* have been tagged as qualifiers:

> In general, adverbs in *-ly* have not been tagged QL even when they serve a qualifying function; they are given the adverb tag RB. There are, however, three exceptions - *awfully, fairly*, and *really*.
>
> (Francis and Kucera, 1979: section 4)

Unfortunately, this statement is not correct. If it was, Brown would be like LOB, not like ICE. In fact, there are more than 290 different word types ending in *-ly*, which are

tagged as qualifiers in the Brown Corpus. The reason why the Manual contains this statement, which bears no relation to the corpus it is supposed to describe, is not clear.

Again there are numerous inconsistencies, that is, errors, in the tagging. Some examples are given in Table 10. The four words in the table are tagged differently, although only words which occur in the same context (before an adjective tagged JJ which they qualify) have been counted. In general, it would appear that the Brown Corpus has a rather large number of erroneous tags, certainly considerably more than in LOB or the Sampler.

		tagged RB	tagged QL
1	exceedingly	1	5
2	sufficiently	2	23
3	terribly	3	5
4	unusually	5	5

Table 10: Inconsistent taggings in Brown

In a work by the Brown authors, which appeared three years after the publication of the tagged Brown Corpus and the accompanying Manual (now the on-line Manual), a very different statement occurs:

> In general, adverbs in -ly immediately preceding and clearly qualifying an adjective or adverb are commonly tagged QL, rather than the general adverb tag RB. Examples are *exceedingly, sufficiently, terribly, unusually*.
>
> (Francis and Kucera, 1982: 10)

This appears to be a correction to the erroneous statement in the Manual, and perhaps the phrase *in general* and the word *commonly* are an acknowledgement of the many errors present in the corpus in connection with this tag. There are errors even for the four words they have themselves chosen to illustrate their statement. The relevant figures have already been given in Table 10 above. If the RB tags are wrong, there is an error rate of about twenty-two percent. For many other degree adverbs ending in *–ly* the error rate is much higher.

6. Interrogative and relative adverbs

There are altogether nine subclasses of interrogative and relative adverbs in the four corpora, just one in LOB, two each in Brown and ICE, and four in the Sampler. Table 11 shows the subclasses and their type and token totals.

a Wh- general adverbs				
B	WRB	20	4,569	wh- adverb
L	WRB	15	5,076	WH-adverb
S	RRQ	14	2,058	wh- general adverb
I	ADV(wh)	16	2,440	wh- adverb
b Wh- degree adverbs				
B	WQL	2	181	wh- qualifier
S	RGQ	1	229	wh- degree adverb
c Wh-ever general adverbs				
S	RRQV	5	70	wh-ever general adverb
d Wh-ever degree adverbs				
S	RGQV	2	28	wh-ever degree adverb
e Relative adverbs				
I	ADV(rel)	8	954	relative adverb

Table 11: Interrogative and relative adverbs

The way these tags are used in the different corpora can be illustrated by considering the follow seven word types: *when, where, why, how* and *whenever, wherever, however*.

6.1 Classification schemes

Each of the four corpora uses a different classification scheme for these words, as shown below in Tables 12 and 13. The schemes for Brown, LOB, and the Sampler can be grouped into a single table, but that for ICE, which is rather different, requires its own table.

Brown		
WRB	when	whenever
	where	wherever
	why	
	how	however
WQL	how	however

LOB		
WRB	when	whenever
	where	wherever
	why	
	how	however

Sampler			
RRQ	when	whenever	RRQV
	where	wherever	
	why		
	how	however	
RGQ	how	however	RGQV

Table 12: The tagging schemes for B, L, and S

ICE				
ADV(rel)	when	when	whenever	ADV(wh)
	where	where	wherever	
		why		
	how	how	however	

Table 13: The tagging scheme for I

The Brown Corpus gives a special tag to the two words *how* and *however* when they are used as qualifiers, as, for example, before adjectives and adverbs. Two subclasses result.

The LOB scheme is a simplification of the Brown scheme, and places all these words in a single subclass.

The Sampler reverts to the distinction made in Brown, and introduces a new one of its own, the difference between words which do and do not end in *-ever*. This is the most complicated scheme with four subclasses.

Finally, ICE ignores both these distinctions, but introduces a new one of its own, the difference between relative and non-relative words. Just as in Brown, there are two subclasses, but the basis for the distinction is quite different.

7. Adverb or preposition

The primary motivation behind this group of tags is the desire to distinguish adverbs which are part of phrasal verbs. Most of these adverbs have prepositions as grammatical homonyms.

LOB distinguishes a second group of adverbs with prepositional homonyms, but which are not part of phrasal verbs.

And the sampler has a special tag for just one word *about*, when it is followed by the infinitive marker *to*.

Table 14 lists these tags.

a Phrasal adverbs				
B	RP	12	6,039	adverb/particle
L	RP	28	8,700	adverbial particle
S	RP	21	6,873	prepositional adverb, particle
I	ADV(phras)	61	7,211	phrasal adverb
b Adverbs with prepositional homographs				
L	RI	21	571	adverb (homograph of preposition)
c Catenative prepositional adverbs				
S	RPK	1	40	prepositional adverb, catenative

Table 14: Words which can be adverbs or prepositions

8. New subclasses of adverbs

In two corpora, new subclasses of adverbs have been created from words which belong, in the other corpora, to the general adverb (or simply adverb) class. Two such subclasses are found only in the Sampler, and three only in ICE. They are listed in Table 15.

a Post-nominal adverbs				
S	RA	50	502	adverb after nominal head
b Nominal introducer adverbs				
S	REX	11	454	adverb introducing nominal construction
c Focusing adverbs				
I	ADV(excl)	16	3,970	exclusive adverb
I	ADV(partic)	23	964	particularizer adverb
I	ADV(add)	22	3,349	additive adverb

Table 15: New adverb subclasses in the Sampler and ICE

8.1 The two subclasses of the Sampler

The first subclass of the Sampler, with the tag RA, is basically for a small set of about ten adverbs which commonly follow a number or time noun. The most common is *ago*, as, for example, in *years ago*. Four more are the abbreviations *AM* and *PM*, and *BC* and *AD*. Words in which these abbreviations are joined with a preceding time word or number have also been tagged RA, for example *1am*, *2.30pm*, *863BC*. This is the reason for the large number of word types – fifty-one - in the table. Since the number of times and dates which can precede an abbreviation is extremely large, the number of such compound single words may vary greatly, depending on the size and nature of the corpus. The total may become very large indeed.

The second new subclass contains just a few word types, including the abbreviations *e.g.* and *i.e.*, as well as their multiword equivalents *for example* and *that is*. In ICE, members of this class are included in the appositive subclass of a new non-adverb "connective" word class, with the tag CONNEC(appos) (see section 9.3 and 9.3.2).

8.2 The three subclasses of ICE

Some of the adverbs belonging to these classes are listed in the 1972 *A Grammar of Contemporary English* under the heading *Focusing adjuncts*.

> Focusing adjuncts constitute a fairly limited set of items, mostly adverbs, but also some prepositional phrases. Common items are listed below.
> (Quirk *et al*., 1972: 431, and Quirk *et al*., 1985: 604)

Three lists follow: there are ten items in the list of exclusives, twelve in that for particularizers, and thirteen in that for additives.

Almost exactly the same comment, and exactly the same three lists are found again in the 1985 *Comprehensive Grammar of the English Language*. This time however the heading is not *Focusing adjuncts* but *Focusing subjuncts*.

In 1996, Greenbaum comments about these three ICE subclasses, together with the relative subclass:

> we were inclined to include subclasses that comprised a limited number of lexical items that we could list in full
>
> (Greenbaum, 1996b: 95)

However, other closed classes listed in the same grammar are not present in ICE.

9. New word classes for words once classed as adverbs

By the time the tagging of the Brown Corpus was first being planned, grammarians were already removing a few particularly important words from the adverb class, and setting up newly-named one-member word classes. In 1972, *A Grammar of Contemporary English* contained the following statement:

> Because of its great heterogeneity ... some grammarians have removed certain types of items from the class entirely and established several additional classes rather than retain these as subsets within a single adverb class.
>
> (Quirk *et al.*, 1972: 267, and Quirk *et al*., 1985: 436)

The two most important instances are undoubtedly the word *not*, and the word *there*, as it is used, for example, in sentences which begin with *There* and some form of the verb *BE*. In such contexts, it is now usually termed the "existential" *there*.

Two new word classes, connectives and reaction signals, the first with two subclasses, are found only in ICE.

9.1 Negative *not* and *n't*

Table 16 lists the relevant tags and frequencies. The absence of such a tag in ICE is immediately apparent.

B	*	3	4,615	not, n't
L	XNOT	8	7,454	not, n't
S	XX	5	6,140	not, n't

Table 16: Negative *not*

The Brown tag is an asterisk, the only tag in the first three corpora not to contain at least one letter. In ICE, there is also an UNCLEAR tag which is a question mark.

9.1.1 The tagging of the first three corpora

Brown, LOB and the Sampler all treat *not* and its variant as a word class with just one member. Brown however constructs combined tags for orthographic words ending in *n't.* Combined tags in Brown and, in particular, contracted negatives, are further discussed in sections 12 and 12.1.

9.1.2 The tagging of ICE

ICE, however, treats the word *not* as an adverb. This is particularly unexpected, since the first director of the ICE project, Sidney Greenbaum had always endorsed the treatment of *not* as a separate word class:

> there are some words that do not fit anywhere and should be treated individually, such as the negative *not*
>
> (Greenbaum, 1991: 69-70)

> Some words do not fit well into any of the classes. Among them are: ... the negative particle *not*
>
> (Greenbaum, 1995: 93)

Although in ICE the negative word *not* is not distinguished in any way from any other general adverb, its contracted form *n't* is always represented by the tag feature *neg*. The word *can't*, for example, is tagged *AUX(modal,pres,neg).* Thus, ICE has two quite different ways of tagging two variants of the same lexeme, which does not seem to be a very good idea. All the other corpora, on the other hand, have a single representation for both variants.

But this is not the end of the story. The negative feature *neg* is not reserved for use with *n't*, but is also used with a small group of negative words. The following words are the only words other than words ending in *n't* assigned this feature in ICE: *no, none, neither, nobody, no one, no-one, nothing, nowhere.* (That statement is not quite true because some verbs and auxiliaries with no final *n't* are wrongly tagged *neg*). The words in this small group are always tagged as pronouns.

In section 13 there is a table for words with the negative feature in ICE.

9.2 Existential *there*

Existential *there*, on the other hand, is treated as a separate one-member class in all four corpora.

B	EX	3	2,169	existential there
L	EX	3	2,793	existential there
S	EX	3	2,215	existential there
I	EXTHERE	1	3,444	existential there

Table 17: Existential *there*

9.3 Connectives and reaction signals

Three new tags have been created in ICE, none of which have any equivalents in the other corpora.

The tags and frequencies are found in Table 18.

I	CONNEC(ge)	198	18,944	general connective
I	CONNEC(appos)	42	1,666	appositive connective
I	REACT	126	10,911	reaction signal

Table 18: Connectives and reaction signals

The first new tag class, called connectives, has two subclasses: general connectives and appositive connectives. The second new class is that of reaction signals.

Both the general connectives and the reaction signals are used for many words which are tagged as adverbs in all the other corpora, But they are also used for words which in some or all of the other corpora are not so tagged. The appositive connectives, on the other hand, are mostly tagged as adverbs in the other corpora.

9.3.1 General connectives

Common general connectives are *however, now, so, then* and *therefore*, and all of them have grammatical homonyms in other adverb subclasses.

A number of words in this class also have grammatical homonyms in non-adverb classes.

9.3.2 Appositive connectives

The members of this subclass are broadly identical to those placed in the new adverb subclass, REX, in the Sampler (see section 8.1).

9.3.3 Reaction signals

Common reaction signals are *absolutely, certainly, definitely*, and *indeed*.

A number of words in this class have grammatical homonyms in other adverb subclasses.

10. Inflected adverbs

All English grammars discuss the comparative and superlative forms of adverbs. And all four corpora assign distinctive tags to these words. Two corpora, Brown and LOB, consider all such forms to be inflected forms of manner adverbs, but the other two treat a small number as inflected forms of degree adverbs. There is an additional comparative additive adverb in ICE.

Traditionally the plural and genitive inflections are associated with nouns and pronouns. In Brown and LOB, special plural and genitive tags are therefore reasonably assigned to certain words in the ambiguous adverbial noun class - particularly those which are later tagged as some form of noun in the Sampler and ICE - forms such as *Mondays* and *Monday's*,

However, genitive endings have always posed a problem for the writers of English grammars. The reason is that a genitive ending may be added to almost any noun phrase in English, and a noun phrase may end in almost any word at all. One solution is to treat the genitive ending as a phrasal affix rather than the inflection of a word, and to tag it with its own special tag, as if it were a word like any other. Both the Sampler and ICE do this, but in Brown and LOB, a new genitive tag is created for any adverb ending in a genitive inflection, and indeed for other word classes too. As it happens, the only adverb in both Brown and LOB with such an inflection is the word *else's*, so this word is tagged as if it were a genitively inflected manner sdverb. This word, however, is the thin end of the wedge since not just adverbs but words from any word class may occur in some future corpus followed by a genitive ending. Tagging the genitive ending as if it were a separate word avoids a potential proliferation of tags, and the theoretical problem of whether the genitive ending is an inflection in some cases but a phrasal affix in others.

10.1.1 The comparative and superlative tags of manner adverbs

a Comparative inflections				
B	RBR	26	1,187	comparative adverb
L	RBR	25	1,357	comparative adverb
S	RRR	25	1,107	comparative general adverb
I	ADV(ge,comp)	24	826	comparative general adverb
b Superlative inflections				
B	RBT	12	101	superlative adverb
L	RBT	12	103	superlative adverb
S	RRT	16	109	superlative general adverb
I	ADV(ge,sup)	10	49	superlative general adverb

Table 19: Adverbial inflections of manner adverbs

10.1.2 The comparative and superlative tags of degree adverbs

a Comparative inflections				
S	RGR	2	1,032	comparative degree adverb
I	ADV(inten,comp)	14	1,291	comparative intensifier adverb
b Superlative inflections				
S	RGT	2	589	superlative degree adverb
I	ADV(inten,sup)	8	545	superlative intensifier adverb

Table 20: Adverbial inflections of degree adverbs

The Sampler has selected just four words for this tag, the words *more* and *less*, and *most* and *least*. ICE has not so strictly limited the words it selects.

10.1.3 The comparative tag of an additive adverb

In ICE the comparative feature has been assigned three times to the one word in the additive subclass, *further*. Since this word is tagged as the comparative of a general adverb in many other instances, either these three tags, or the much larger number of incompatible tags are errors. Table 21 is included, since, in principle, such tags seem possible.

I	ADV(add,comp)	1	3	comparative additive adverb

Table 21: The comparative inflection of another adverbial subclass

10.2.1 Plural nominal inflections of adverbial nouns

The plurals of days of the week are commonest examples, and are found in both corpora.

B	NRS	6	17	plural adverbial noun
L	NRS	8	84	plural adverbial noun

Table 22: Plural inflections

10.2.2 Genitive nominal inflections of adverbs

a Of manner adverbs				
B	RB$	1	9	-
L	RB$	1	6	adverb + genitive
b Of time and place adverbs				
B	NR$	13	77	possessive adverbial noun
L	NR$	13	48	singular adverbial noun + genitive

Table 23: Genitive inflections

The first tag in the table has no descriptive name because it is not present in the tag list of the online Manual, where other such names are to be found.

The only word tagged RB$ is the word *else's* discussed in 10.1 above. The tag NR$ is commonly found with days of the week.

11. Multiwords

The first multiwords in an electronic corpus make their appearance in LOB. In Brown, which preceded LOB, there are no multiwords. In the Sampler, the number increases considerably, and in ICE the number again increases, this time dramatically.

Multiwords are valued for several reasons: they remove the problems associated with tagging difficult groups of words, such as *as yet* or *in full*, and they result in more successful automatic tagging and parsing.

11.1 Multiwords in LOB

In LOB, the constituent words of a multiword are assigned the same tag, but all tags except the first are marked by a following neutral double quotation mark - called a ditto mark - to show that they are part of a multiword. Such tags are conveniently called ditto tags. For example, *by the by* is tagged *by_RB the_RB" by_RB"*.

In LOB, all adverb multiwords are manner adverbs. Table 24 shows the frequency of these items.

L	RB"	61	1,785	adverb ditto tag

Table 24: The adverb ditto tag in LOB

This method of tagging has disadvantages when word lists are constructed. There is nothing to show that the first constituent is part of a multiword, and in addition it has a tag belonging to the multiword as a whole rather than one inherent to itself. When lists are constructed, however, these first constituents of multiwords form part of the same list as independent words. Thus the first word of *by the by* will form part of a list of LOB manner adverbs, as if it too was a manner adverb. In fact, there is no word *by* in LOB tagged as a manner adverb, apart from words which are the first words of multiwords. This may be misleading, unless such lists are properly interpreted..

11.2 Multiwords in the Sampler

In the Sampler, multiwords are tagged as wholes, and the parts are not assigned independent tags, as they are in LOB.

11.3 Multiwords in ICE

In ICE, multiwords are tagged as wholes in data files, just as in the Sampler.

But the accompanying HELP facility - in effect the Manual for ICE - displays them differently: here the constituent words of multiwords are tagged independently, as in LOB, but each constituent is followed by two numbers separated by a slash: the first is the number of the constituent, the second, the total number of constituents in the multiword. This is

different from LOB, where the first constituent is not marked. For example *by and large* would be tagged *by_ADV(ge)1/3 and_ADV(ge)2/3 large_ADV(ge)3/3*.

Moreover, ICECUP, the accompanying facility for displaying tagged and parsed sentences, also displays each word of a multiword with its own tag, although without the following pair of numbers. In the ICECUP display all the consituents of a multiword are linked together by yellow lines.

12. Combined tags in Brown and ICE

In the Brown Corpus single orthographic words are regularly given a single tag. This is also true of words like *there's* and *isn't* which have two-word othographic variants. However, the single tag is made up of the two tags of the words of the variant. Contracted negatives simply add the negative tag, an asterisk, directly to the tag for the first word. All other such words separate the parts of the tag with a plus sign.

In LOB and the Sampler, all these words are treated differently: they are split into two separate orthographic words, and each such word is given its own tag.

In ICE contracted negatives are left as parts of single orthographic words, just as in Brown, but the tags for such words include the negative feature *neg*. One tag for *isn't*, for example, is *V(cop,pres,neg).* Other words of this kind, such as *there's*, which do not have contracted negatives are treated as in LOB and the Sampler.

12.1 Contracted negatives in Brown

In Brown, single orthographic words like *isn't* ending in a contracted negative, and which have two-word variants – in this case *is not* – are tagged with a single tag made up of the tags of the two-word variant. The two word variant is tagged *is_BEZ not_**, so the contracted word is tagged *isn't_BEZ**. The word *cannot* is tagged similarly as MD*.

The contracted negative forms which occur in Brown are listed below in table 25, together with illustrative words.

	Tag	Type total	Token total	Example
1	BED*	1	22	were'nt
2	BEDZ*	1	155	wasn't
3	BEM*	1	9	aint't
4	BER*	2	48	aren't
5	BEZ*	3	117	isn't
6	DO*	2	488	don't
7	DOD*	2	405	didn't
8	DOZ*	2	90	doesn't
9	HV*	2	42	haven't
10	HVD*	1	100	hadn't
11	HVZ*	2	22	hasn't
12	MD*	11	867	couldn't

Table 25: Contracted negative tags in Brown

Adding the token totals in this table gives the total number of tokens for the contracted negative *n't*, which is 2,365. This total, however, also includes the total for the *not* when it is part of the word *cannot*.

12.2 Adverbs with combined tags in Brown

In the Brown Corpus, single orthographic words like *there's* which have two-word variants – in this case *there is* – are tagged with a single tag made up of the tags of the two-word variant joined by a plus sign. The two word variant is tagged *there_EX is_BEZ*, so the contracted word is tagged *there's_EX+BEZ*.

Six Brown adverb tags have combined tags of this kind. They are listed below in table 26 with illustrative words.

	Tag	Type total	Token total	Example	Relevant table
a	adverb				
1	RB+BEZ	2	13	here's	4
2	RB+CS	2	3	soon's	
b	adverbial noun				
1	NR+MD	1	1	today'll	6
c	wh-adverb				
1	WRB+BER	1	1	where're	11
2	WRB+BEZ	2	14	where's	
3	WRB+DO	1	1	howda	
4	WRB+DOD	2	6	how'd	
5	WRB+DOD*	1	1	whyn't	
6	WRB+DOZ	1	1	how's	
8	WRB+IN	1	1	why'n	
7	WRB+MD	1	1	where'd	
d	adverb/particle				
1	RP+IN	2	4	outta	14
e	existential *there*				
1	EX+BEZ	1	105	there's	17
2	EX+HVD	1	3	there'd	
3	EX+HVZ	1	2	there's	
4	EX+MD	2	4	there'd	
f	comparative adverb				
1	RBR+CS	1	1	more'n	19

Table 26: Combined tags in Brown

These figures can be used to make any desired adjustment to the various token totals in this paper, if more exact comparisons are needed.

13. Words with the negative feature in ICE

In section 9.1.2 we have already discussed the use of the negative feature in ICE, and how it is used for both negative pronouns and for words ending in *n't*. This latter group consists of verbs and auxiliaries. Because the subcategorization of verbs and auxiliaries using features is very complicated in ICE, a complete list of tags with the negative feature is not given here, but rather a list of the three primary classes to which the negative feature can be appended, together with token totals.

	Class of tag	Type total	Token total
1	PRON	33	1,754
2	V	23	841
3	AUX	83	5,349

Table 27: Tags with the negative feature in ICE

Tags with the class label V or AUX, together with the negative feature *neg*, are used for words ending in the contracted negative *n't* (and for *cannot*), and adding the last two token totals on the right gives the token total for this form – 6,190. Again, this total includes the count for the *not* of *cannot*.

14. Adverbs with a discontinuous tag in ICE

In ICE, a new kind of tag is introduced for discontinuous constituents of a word or multiword. The parts of a discontinuous item are marked by a special discontinuous feature, abbreviated *disc*. In such cases, the parts of a discontinuous item are tagged separately in the data files, and each tag is modified by the addition of this feature followed by an identifier digit. In the spoken part of the corpus, the single-word adverb *overboard* and the multiword adverb *a bit* are spoken with a pause in the middle of the word. When these words are transcribed, a PAUSE tag is used in the places where the pauses occur, so that the two words now have two discontinuous parts, each requiring its own tag. Thus, they are transcribed as *over PAUSE board* and *a PAUSE bit.* Each part of the first adverb has the tag *ADV(ge,disc1)*, and each part of the second the tag *ADV(inten,disc1)*. This discontinuous feature can be used for the tagging of any kind of word or multiword if it is necessary. It is only found in a few instances with adverbs, however. Table 28 shows the four tags found.

1	ADV(ge,disc1)	2	2	discontinuous general adverb
2	ADV(inten,disc1)	4	4	discontinuous intensifier adverb
3	CONNEC(ge,disc1)	3	3	discontinuous general connective
4	ADV(inten,comp,disc1)	1	2	discontinuous comparative intensifier adverb

Table 28: Discontinuous adverb tags

The markup itself does not allow the parts of a discontinuous word to be distinguished from those of a discontinuous multiword, as can be seen from the two examples above. It is not possible to tell from the markup whether the two constituents *over*

and *board* are part of a single word *overboard* or of a, in this case, most unlikely multiword *over board*.

15. The feature *ignore* in ICE

A feature *ignore* can be added to any tag, and can be seen on the ICECUP display as a crossed (out) item. The editors have marked certain items with this feature, particularly in transcriptions of spoken English, so that a more regularized and thus more easily parsable sentence results. Repetitions and interruptions are typically so marked. An example is *I actually_ADV(ge,ignore) actually_ADV(ge) think so*, in which the first occurrence of *actually* is marked with an ignore feature.

The fifteen adverb tags with this feature are listed in table 29 below, together with their frequencies.

In the whole corpus (excluding punctuation and pause tags) there are 3,425 different tagged word types and 24,398 tokens with the *ignore* feature. Table 29 lists all the adverb tags with this feature: there are 235 types, and 1,666 tokens.

1	ADV(ge,ignore)	100	415
2	ADV(inten,ignore)	47	221
3	ADV(wh,ignore)	4	134
4	ADV(rel,ignore)	2	15
5	ADV(phras,ignore)	14	49
6	ADV(excl,ignore)	4	123
7	ADV(partic,ignore)	4	9
8	ADV(add,ignore)	5	19
9	EXTHERE(ignore)	1	168
10	CONNEC(ge,ignore)	33	407
11	CONNEC(appos,ignore)	9	18
12	REACT(ignore)	7	58
13	ADV(ge,comp,ignore)	2	3
14	ADV(inten,comp,ignore)	2	23
15	ADV(inten,sup,ignore)	1	4

Table 29: Ignored adverb tags

16. Anomalous adverb tags in ICE

Some tags lack certain, apparently obligatory, features. For the present, such tags will be called "anomalous" tags. An examination of the accompanying parse tree often shows that some of these tags are certainly erroneous: this is clear when the tag does not have a feature or features which have nevertheless been copied up to the adverb phrase. Such featureless tags are also not mentioned in the Help facility, previously the corpus Manual. Table 30 lists the five relevant adverb tags. We have not yet determined whether all tokens tagged in this way are errors.

1	ADV	23	50	(anomalous) adverb
2	CONNEC	24	389	(anomalous) connective
3	CONNEC(ignore)	3	7	ignored (anomalous) connective
4	ADV(comp)	7	10	anomalous comparative
5	ADV(sup)	1	1	anomalous superlative

Table 30: Anomalous adverb tags

17. Conclusions

In 1982, the authors of the Brown Corpus gave the following explanation for their choice of tags:

> Since the primary purpose of tagging the Brown Corpus was to facilitate automatic or semiautomatic syntactic analysis, the rationale for the tagging is basically syntactic, though some morphological distinctions with little syntactic significance have been made.
> (Francis and Kucera, 1982: 9)

We do not know of any detailed discussion or exemplification of this statement by the authors, nor do they appear to have attempted themselves any such automatic syntactic analysis. Some tags certainly appear to have such a rationale, and are used for the first time in the Brown Corpus. For example, the ambiguous nominal adverb, adverbial noun, and adverb/particle (probably) had their origin in the Brown Corpus. The adverb/particle tag is designed to allow easier identification of phrasal verbs, but (as far as we know) there are no parsed corpora which treat phrasal verbs as a syntactic unit. The parsed ICE Corpus identifies certain adverbs as phrasal adverbs, but they do not belong with the verb they accompany in the same syntactic constituent, nor is the verb which accompanies the phrasal adverb in any way marked as a phrasal verb. Nor is it clear how syntactic analysis can be helped by such tags as adverbial nouns and nominal adverbs. In the Sampler Corpus they find no place, even though skeleton parsing was being undertaken at Lancaster at the same time as the tagging of the BNC was taking place. Even the recognition of the negative adverb *not*, which is perhaps another innovation of the Brown Corpus, does not seem to be

required for automatic parsing: for instance, this tag is not found in the ICE Corpus. And the ICE negative feature *neg* does not appear to play any role in parsing.

The tagset of the LOB Corpus is based directly on that of Brown. While planning and executing the tagged LOB Corpus, one of the main interests of the authors undoubtedly became the possibility of tagging texts automatically with a minimum of errors. The fact that the automatic tagger used frequency data based on the Brown Corpus ensured that the LOB tagset remained close to that used for Brown. Nevertheless, a number of changes were made. For example, the qualifier tag class was drastically reduced from several hundred word types to less than twenty. As a result fewer errors were made by the automatic tagger, and the tag acquired a much clearer significance. Another change which was made to improve the automatic tagger completely altered the tagging of all the corpora which followed LOB, and as a result of its presence in the LOB Corpus probably greatly influenced dictionary and grammar book writers who came after: this change is the introduction of multiwords. In the following ICE corpus, it is clear that the use of multiwords not only can improve the accuracy of automatic tagging, but also can remove difficulties for automatic parsing programs. But the best change of all was a tremendously improved accuracy and consistency in the tagging compared with the Brown Corpus. The LOB Corpus is in every way a superior achievement in this respect.

The Sampler Corpus introduces a number of new subclasses and dispenses with a number of old ones. Basically, however, it is a continuation of the Brown-LOB tagset. Again there is little published discussion that we are aware of as to why these changes are introduced, or explanations of whether, and if so how, they effect automatic parsing. Are new (for tagged corpora that is) tags like those for time and place actually semantically-based classes introduced to tidy up the messy adverbial noun and nominal adverb tags without actually discarding them, or are they tag classes with some importance for the automatic parsing programs that were being developed when this tagset was being evolved? Like LOB, however, the tagging of adverbs in the Sampler seems to be of a very high standard indeed.

Finally ICE. The tagset here is a strange mixture. Some new tags are apparently entirely semantically-based - the exclusive, particularizer, and additive tagsets - but some apparently similar old tags have been discarded – namely the time and place tags. Perhaps the reason for the introduction of these tags is the spoken nature of much of the corpus, but they do not seem to have any significance for automatic parsing. Some tags on the other hand seem very desirable: surely everyone wants to know where the relative adverbs occur. The relative adverb label is used in every grammar book but only finds exemplification in a corpus in ICE. The tagging is not as good as LOB or the Sampler: nevertheless this is a very small criticism indeed when it is considered that a wonderful resource for exemplifying and studying syntactic patterns in English has been created in the ICE Corpus.

We have greatly enjoyed working with these corpora. We hope that you will use them and enjoy them as much as we have.

16. Synoptic table of adverb tags

In the synoptic table below, the numbers in the boxes on the left refer to the corresponding tags in the coloured boxes on the right. Thus number *1* refers to the tag *RB* coloured yellow on the right.

Each group of related tags is separated from other groups by a thick black line. The name used in this paper for each set of tags is given in italics, followed by the number of the section where they are discussed. On the far right is the number of the table where their frequencies and descriptive names can be found.

Note that the Sampler tag *REX* and the ICE tag *CONNEC(appos)* are used for essentially the same items (see sections 8.1 and 9.3.2). For this reason they have been entered twice in the table, in the appropriate boxes. The second entries are italicised and uncoloured. These are the only examples of equivalent tags occurring in different sets.

Combined tags in Brown (tables 25 and 26), and tags for verbs and auxiliaries in ICE which have a negative feature representing *n't* have not been included. Only one example each is given for discontinuous, ignored, and anomalous tags. Otherwise the table is complete.

1	2	3	4	*Manner (3)*	RB	RB	RR	ADV(ge)	*4*
5	6			*Time & place (4)*	RN	RN			*6*
7	8				NR	NR			
		9					RT		
		10					RL		
11	12	13	14	*Degree (5)*	QL	QL	RG	ADV(inten)	*9*
15	16				QLP	QLP			
17	18	19	20	*Interrogative and*	WRB	WRB	RRQ	ADV(wh)	*11*
21		22		*relative (6)*	WQL		RGQ		
		23					RRQV		
		24					RGQV		
			25					ADV(rel)	
26	27	28	29	*Adverb or*	RP	RP	RP	ADV(phras)	*14*
	30			*preposition (7)*		RI			
		31					RPK		
		32		*New subclasses (8)*			RA		*15*
		33	*45*				REX	*CONNEC(appos)*	
			34					ADV(excl)	
			35					ADV(partic)	
			36					ADV(add)	
37	38	39		*New classes (9)*	*	XNOT	XX		*16*
40	41	42	43		EX	EX	EX	EXTHERE	*17*
			44					CONNEC(ge)	*18*
		33	45				*REX*	CONNEC(appos)	
			46					REACT	
47	48	49	50	*Inflections (10)*	RBR	RBR	RRR	ADV(ge,comp)	*19*
51	52	53	54	*Comp. and sup. manner*	RBT	RBT	RRT	ADV(ge,sup)	
		55	56	*Comp. and sup. degree*			RGR	ADV(inten,comp)	*20*
		57	58				RGT	ADV(inten,sup)	
			59	*Comp. additive*				ADV(add,comp)	*21*
60	61			*Plural advl noun*	NRS	NRS			*22*
62	63			*Genitive manner*	RB$	RB$			*23*
64	65			*Genitive advl noun*	NR$	NR$			
	66			*Ditto manner (11)*		RB”			*24*
	67			*Disc. manner (14)*				ADV(ge,disc1)	*28*
	68			*Ignored manner (15)*				ADV(ge,ignore)	*29*
	69			*Anomalous (16)*				ADV	*30*

Table 31: Synoptic table

References

Atwell, E., G. Demetriou, J. Hughes, A. Schriffin, C. Souter, S. Wilcock. (2000)
A comparative evaluation of modern English corpus grammatical annotation schemes. ICAME Journal, vol. 24, pp. 7–23.

BNC Baby CD-ROM Release 2 (2005) Oxford: The Humanities Computing Unit of Oxford University (contains Brown and the Sampler).

Francis, W. N. and H. Kucera (1979) Manual of Information. Providence, Rhode Island: Department of Linguistics, Brown University. Available from http://icame.uib.no/brown/bcm.html

Francis, W. N. and H. Kucera (1982) Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin.

Greenbaum, S. (1991) An Introduction to English Grammar. Harlow: Longman.

Greenbaum, S. (1996a) The Oxford English Grammar. Oxford: Oxford University Press.

Greenbaum, S and Ni Yibin (1996b) About the ICE Tagset, in Greenbaum (ed.) Comparing English Worldwide, pp. 93–109. Oxford: Clarendon Press.

Greenbaum, S. (ed.) (1996c) Comparing English Worldwide: The International Corpus of English. Oxford: Clarendon Press.

ICAME CD-ROM Version 2 (1999) Norway: Aksis, University of Bergen (contains Brown and LOB).

ICE-GB CD-ROM Release 2 (2006) London: Survey of English Usage, Department of English, University of London.

Johansson, S., E. Atwell, R. Garside and G. Leech (1986) The Tagged LOB Corpus: User's Manual. Available on-line from www.comp.lancs.ac.uk/ucrel/local/lob/

Jurafsky, D. and J. Martin. (2000) Speech and Language Processing. New Jersey: Prentice-Hall.

Leech, G. (1997) BNC Sampler Corpus: Guidelines to Wordclass Tagging. Available on-line from www.comp.lancs.ac.uk/ucrel/bnc2sampler/guide_c7.htm

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1972) A Grammar of Contemporary English. London: Longman.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985) A Comprehensive Grammar of the English Language. London: Longman.

Thomson, A. J. and A. V. Martinet (1969) A Practical English Grammar. Oxford: Oxford University Press.

van Halteren, H. (ed.) (1999) Syntactic Wordclass Tagging. Dordrecht: Kluwer.