Unit 15 Contrastive and diachronic studies

15.1 Introduction

As noted in units 10.4 – 10.7, corpora are well suited to comparative and diachronic studies. Of the various types of corpora introduced in unit 7, comparable and parallel corpora are particularly useful in contrastive and translation studies (see unit 5). Likewise, diachronic studies have always, in a sense, been corpus-based (cf. also Bauer 2002: 109). This unit presents four excerpts from published material to demonstrate the use of corpora in these two types of language studies. The first two excerpts are concerned with contrastive analysis while the latter two explore language change.

15.2 Altenberg and Granger (2002)

This excerpt discusses the use of comparable and parallel corpora (or 'translation corpora' in the authors' term) in contrastive linguistics (CL), focusing on lexis. Readers are reminded that their proposal to base cross-linguistic contrast on parallel corpora is potentially problematic, as noted in unit 10.6.

Altenberg, B. and Granger, S. 2002. 'Recent trends in cross-linguistic lexical studies' in B. Altenberg, and S. Granger (eds.) *Lexis in Contrast,* pp. 3-48. Amsterdam: John Benjamins.

3. Theoretical and methodological issues

3.1 Some contrastive approaches

Traditionally, CL has been described as involving three methodological steps: description, juxtaposition and comparison (see e.g. Krzeszowski 1990:35). The description includes selection of the items to be compared and a preliminary characterisation of these in terms of some language-independent theoretical model. The juxtaposition involves a search for, and identification of, cross-linguistic equivalents. In the comparison proper the degree and type of correspondence between the compared items are specified.

Modern lexical CL often follows this procedure, but a characteristic feature of recent corpus-based contrastive work is the great variety of approaches employed. This is largely due to the expansion of the field and the new research possibilities that multilingual corpora and search tools offer. The methodology chosen and the delicacy of the analysis depend to a large extent on the purpose of the analysis, e.g. whether it is primarily 'theoretical' (focusing on a contrastive description of the languages involved) or 'practical' (intended to serve the needs of a particular application). This in turn may determine the role that the corpus is allowed to play in the analysis. One distinction that is sometimes made in corpus linguistics, and which is also applicable to CL, is that between 'corpus-based' and 'corpus-driven' approaches (see e.g. Francis 1993 and Tognini Bonelli 2001 and in this volume). The former may involve any work – theory-driven or data-driven – that makes use of a corpus for language description, but it is also used in a restricted sense to refer to studies which start from a model postulating a cross-linguistic difference or similarity on theoretical grounds and use a multilingual corpus to confirm, refute or enrich the theory. The latter approach, on the other hand, may start from an implicit or loosely formulated assumption but uses the corpus primarily to discover types and degrees of cross-linguistic correspondence and to arrive at theoretical statements. In practice, however, the distinction may be slight. The difference lies rather in the importance attached to the initial assumptions and the role that the data play in the analysis. Here we shall use the term 'corpus-based' as an umbrella term covering both types of corpus-informed studies.

In the following sections we shall briefly examine some of the theoretical and methodological issues involved and how these have been approached in some recent corpus-based contrastive studies of lexis.

3.2 Tertium comparationis and translation equivalence

Any cross-linguistic comparison presupposes that the compared items are in some sense similar or comparable. That is, to be able to say that certain categories in two languages are similar or different it is necessary that they have some common ground, or *tertium comparationis*. For lexis it is obvious that the compared items should express 'the same thing', i.e. have the same (or at least similar) meaning and pragmatic function (see James 1980:90f.). However, what exactly this 'thing' is is not always obvious, and the problem of identifying *a tertium comparationis* in CL has been discussed a great deal in the past (see e.g. James 1980:169ff., Krzeszowski 1990, and Chesterman 1998:27ff.).

Krzeszowski (1990:23f) has distinguished seven types of equivalence: statistical equivalence, translation equivalence, system equivalence, semanticosyntactic equivalence, rule equivalence, substantive equivalence and pragmatic equivalence. However, although there is something to say for this taxonomic approach, it seems that the only way we can be sure that we are comparing like with like is to rely on translation equivalence (see James 1980: 178). Chesterman (1998:37ff.) develops this in the following way. Any notion of equivalence is a matter of judgement. Similarly, cross-linguistic equivalence is not absolute, but a matter of judgement or, more precisely, translation competence. "On this view, estimations of any kind of equivalence that involves meaning must be based on translation competence, precisely because such estimations require the ability to move *between* utterances in different languages. Translation competence, after all, involves the ability to *relate* two things" (ibid.: 39).

The fact that equivalence is a relative concept also has another consequence. It is not realistic to proceed from a *tertium comparationis* that is based on 'identity of meaning'. For one thing, this would be putting the cart before the horse and we would run the risk of methodological circularity: the result of the contrastive analysis would be no more than the initial assumption (cf. Krzeszowski 1990:20). For another, the area we want to explore is often fuzzy and impossible to define satisfactorily (e.g. epistemic modality or pragmatic particles). In such cases we cannot start from a *tertium comparationis* that is founded on equivalence in a strict sense (identity of meaning). Instead, what we have to do – and what we generally do – is to start from a perceived or assumed similarity between cross-linguistic items (cf. James 1980: 168f.). Viewed in this way, CL becomes a way of refining initial assumptions of similarity. Chesterman (1998:58) expresses this as follows:

> In this methodology, the *tertium comparationis* is thus what we aim to arrive at, after a rigorous analysis; it crystallizes whatever is (to some extent) common to X and Y. It is thus an explicit specification of the initial comparability criterion, but it is not identical with it – hence there is no circularity here. Using an economic metaphor, we could say that the *tertium comparationis* thus arrived at adds value to the initial perception of comparability, in that the analysis has added explicitness, precision, perhaps formalization; it may also have provided added information, added insights, added perception.

The crucial role that translation equivalence plays in CL has important methodological consequences. We have already described the differences between comparable corpora and translation corpora (Section 2.1). When items are compared across comparable corpora, it is difficult to know if we are comparing like with like. Any judgement about cross-linguistic equivalence (or similarity) must be based on the researcher's translation competence: This is true at both ends of the analysis: initially, when items are selected for comparison, and finally, when the results of the comparison are evaluated. When we use translation corpora the situation is different. Although we normally start with an initial assumption about cross-linguistic similarity – the very basis for comparing anything at all – we can place more reliance on the translations found in the corpus. The corpus can be said to lend an element of empirical inter-subjectivity to the concept of equivalence, especially if the corpus represents a variety of translators.

However, despite the usefulness of translation corpora, to what extent can we trust the translations we find in them? Can we treat all the translations that turn up as cross-linguistic equivalents? There does not seem to be a simple answer to this question. In one sense, every translation is worth considering as a potential translation equivalent as it reflects the translator's 'competence': However, translations are rarely literal renderings of the original. Translators transfer texts from one language (and culture) to another and the translation therefore tends to deviate in various ways from the original. We have already mentioned possible translation effects – traces of the source language or universal translation strategies – and they may involve additions, omissions and various kinds of 'free' renderings that are either uncalled for or motivated by cultural and communicative considerations.

How, then, can we determine which translations should be regarded as 'equivalents' in a stricter sense? One solution has been to resort to the procedure of 'back-translation' (see Ivir 1983, 1987), i.e. to restrict the comparison to forms in L2 that can be translated back into the original forms in L1. This is likely to eliminate irrelevant differences that are due to the translator's idiosyncrasies or motivated by particular communicative or textual strategies.

Another solution is to rely on recurrent translation patterns, i.e. to resort to a quantitative notion of translation equivalence (cf. Kzreszowski 1990:27). If several translators have used the same translation, this obviously increases its relevance. However, this too implies a risk: by restricting the comparison to recurrent translations we may throw away valuable evidence and miss the cross-linguistic insights that 'unexpected' translations often provide.

A variant of this approach which combines Ivir's idea of back-translation and a quantitative notion of equivalence is to calculate what has been called the 'mutual correspondence' (or translatability) of two items in a bidirectional translation corpus (see Altenberg 1999). If an item *x* in language A is always translated by *y* in language B and, conversely, item *y* in language B is always translated by *x* in language A, they will have a mutual correspondence of 100%. If they are never translated by each other their mutual correspondence will be 0%. In other words, the higher the mutual correspondence value is, the greater the equivalence between the compared items is likely to be. Although the mutual correspondence of categories in different languages seldom reaches 100% in a translation corpus (even 80% seems to be a comparatively high value), a statistical measure of translation equivalence can be a valuable diagnostic of the degree of correspondence between items or categories in different languages (see e.g. Altenberg 1999 and Ebeling 1999:257ff.). However, it does not tell us where to draw the line between equivalence and non-equivalence. Ultimately, the notion of equivalence is a matter of judgement, reflecting either the researcher's or the translator's bilingual competence. Both involve a judgement of translation equivalence.

15.3 McEnery, Xiao and Mo (2003)

McEnery, Xiao and Mo (2003) explore aspect marking in English and Chinese, using the FLOB/Frown corpora and a comparable Chinese corpus, the Lancaster Corpus of Mandarin Chinese (LCMC) (see unit 7.4 for a description of the three corpora). The study demonstrates some important similarities and differences in the distribution of aspect markers in Chinese, British English and American English. This excerpt provides background knowledge for case study 6 in Section C and demonstrates how comparable corpora may be used to explore a specific feature cross-linguistically.

McEnery, A., Xiao, Z. and Mo, L. 2003. 'Aspect marking in English and Chinese'. *Literary and Linguistic Computing* 18/4: 361-378.

Having built LCMC, we decided to use the corpus to test a claim made by McEnery and Xiao (2002: 224-5); McEnery and Xiao, based on a study of public health documents in

Chinese and English, claimed that aspect markers occur significantly more frequently in narrative texts than in expository texts. However, McEnery and Xiao only studied one genre. Does this claim hold across a wider range of genres? Also, they only contrasted British English and Chinese. Is the claim true when American English and Chinese are contrasted, or American English and British English? We decided to explore these questions by examining the distribution of aspect markers in the fifteen text categories of the LCMC and FLOB/Frown corpora. In so doing, we were also able to compare the distribution patterns of aspect markers in Chinese and British/American English.

However, before proceeding to the analysis, a brief description of the aspect system of Chinese is needed as Chinese has a very complicated aspect marker system. In Chinese the perfective aspect is marked by *-le*, *-guo*, verb reduplication and resultative verb complements (RVCs) while the imperfective aspect is marked by *zai*, *-zhe*, *-qilai*, and *-xiaqu* (cf. Xiao and McEnery, forthcoming) [later published as Xiao and McEnery (2004b)]. In addition, covert aspect marking is also an important strategy used to express aspectual meanings in Chinese discourse (cf. McEnery and Xiao, 2002: 212). However, as the tagger we used only annotated *-le*, *-guo*, *zai* and *-zhe*, we decided to explore these four aspect markers in LCMC in this study. The frequencies of these aspect markers in LCMC are as shown in Table 4.

Table 4 Distribution of aspect markers in LCMC

Average	Text type	Words (10k)	Frequency	Frequency per 10k words	Percent
Above the average	K	5.8	1674	289	12.00%
	M	1.2	322	268	11.13%
	P	5.8	1384	238	9.88%
	R	1.8	387	215	8.92%
	L	4.8	1024	214	8.88%
	G	15.4	3140	204	8.47%
	N	5.8	1107	191	7.93%
	A	8.8	1539	175	7.26%
Average	Average of frequency per 10k words: 161 (6.68%)				
Below the average	F	8.8	1057	120	4.98%
	C	3.4	365	108	4.48%
	D	3.4	363	106	4.40%
	B	5.4	561	104	4.32%
	J	16.0	1355	84	3.49%
	E	7.6	412	54	2.24%
	H	6.0	231	39	1.62%

English is a less aspectual language with regard to grammatical aspect marking than Chinese. English only differentiates between the simplex viewpoints of the progressive, the perfect and the simple aspect in addition to the complex viewpoint of the perfect progressive (c.f. Biber, Johansson, Leech, Conrad and Finegan, 1999: 461; Svalberg and Chuchu, 1998). In English, perfective meaning is most commonly expressed by the simple past (cf. Brinton, 1988: 52), though the perfect can also mark perfectivity (Dahl, 1999: 34). Imperfective meaning is typically signalled by the progressive, and less often by the perfect progressive. For the purpose of contrasting English aspect marking with Chinese we counted the distribution of the four aspects of English. The frequencies of aspect markers in FLOB and Frown are given in Tables 5-6.

Tables 4-6 show that in both LCMC and FLOB/Frown, the text categories where the frequency of aspect markers is above average (categories L, M, N, P, R, and K) or near to the average (categories A and G) are the five fiction categories plus humour, biography, and press reportage. The text types where aspect markers occur least frequently include reports/official documents, academic prose, skills/trades/hobbies, press reviews, press editorials, religion, and popular lore. In both Chinese and the two major varieties of English considered here, there is a great difference in usage between the first and second groups of texts, which indicates that the two are basically different. Text types like fiction, humour, and biography are narrative whereas reports/official documents, academic prose, and skills/trades/hobbies are expository. Press reportage is a transitory category which is more akin to narrative texts.

Table 5 Distribution of aspect markers in FLOB

Average	Text type	Words (10k)	Frequency	Frequency per 10k words	Percent
Above (or near to) the average	P	5.8	5673	978	11.17%
	L	4.8	4624	963	11.00%
	N	5.8	5255	906	10.34%
	K	5.8	5169	891	10.17%
	M	1.2	997	831	9.49%
	R	1.8	1313	729	8.32%
	A	8.8	5166	587	6.70%
	G	15.4	8257	536	6.12%
Average	Average of frequency per 10k words: 584 (6.67%)				
Below the average	D	3.4	1317	388	4.43%
	F	8.8	3353	381	4.35%
	E	7.6	2724	358	4.09%
	B	5.4	1886	349	3.98%
	H	6.0	1740	290	3.31%
	C	3.4	978	288	3.29%
	J	16.0	4524	283	3.23%

Table 6 Distribution of aspect markers in Frown

Average	Text type	Words (10k)	Frequency	Frequency per 10k words	Percent
Above (or near to) the average	L	4.8	4546	947	10.95%
	M	1.2	1119	933	10.78%
	N	5.8	5349	922	10.66%
	P	5.8	5238	903	10.44%
	R	1.8	1534	852	9.85%
	K	5.8	4815	830	9.59%
	A	8.8	4816	547	6.32%
	G	15.4	7799	506	5.58%
Average	Average of frequency per 10k words: 577 (6.67%)				
Below the average	F	8.8	3397	386	4.46%
	B	5.4	1893	351	4.06%
	E	7.6	2617	344	3.98%
	C	3.4	1155	340	3.93%
	D	3.4	1053	310	3.58%
	J	16.0	4024	252	2.91%
	H	6.0	1368	228	2.64%

Table 7 Distribution of aspect markers in narrative and expository texts

Corpus	Discourse type	Categories	Words	Markers	LL score	Sig. level
LCMC	Narrative	K-R, A, G	494000	10577	2796.53	<0.001
	Expository	B-F, H, J	506000	4344		
FLOB	Narrative	K-R, A, G	494000	36454	7771.37	<0.001
	Expository	B-F, H, J	506000	16522		
Frown	Narrative	K-R, A, G	494000	35216	7950.98	<0.001
	Expository	B-F, H, J	506000	15507		

Log-likelihood (LL) tests indicate that in both Chinese and the two varieties of English, the differences between the distribution of aspect markers in narrative and expository texts are statistically significant (see Table 7). In all of the three corpora, aspect markers occur in narrative texts twice as frequently as in expository texts (2.43 times in LCMC, 2.21 times in

FLOB, and 2.27 times in Frown), which means that the higher frequency of aspect markers in narrative texts over expository texts is a common feature of Chinese and the two major varieties of English.

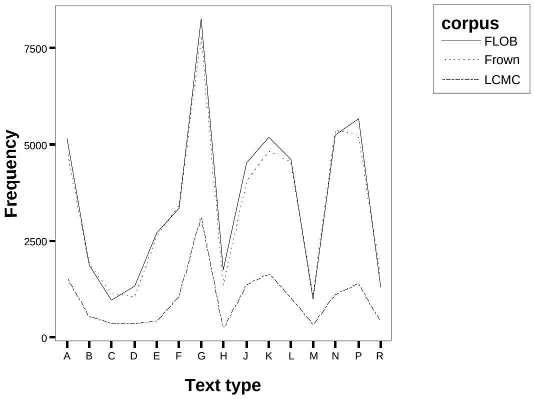

Fig. 1 Distribution of aspect markers (frequency)

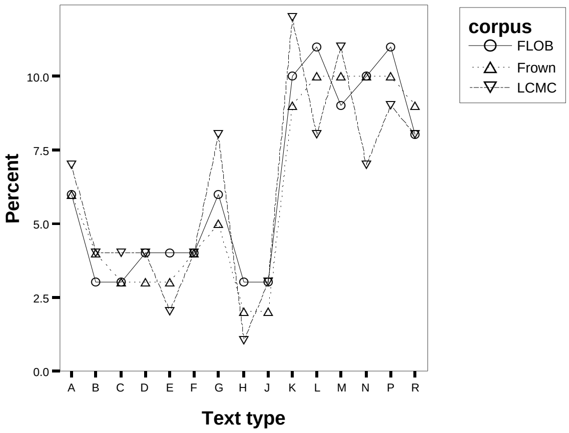

Fig. 2 Distribution of aspect markers (percentage)

These findings confirm those of McEnery and Xiao (2002) and allow us to generalize this claim from the domain studied by McEnery and Xiao, public health, to English/Chinese in general. As can be seen from Fig. 1, while the two languages differ typologically, they show a strikingly similar distribution pattern of aspect markers. It is also interesting to note that while British English and American English have developed variations in spelling (e.g. *behaviour* vs. *behavior*), word choice (e.g. *petrol* vs. *gasoline*), and grammar (e.g. American English has two participle forms for the verb *get*, namely *got* and *gotten* whereas British English only uses the form *got*) (cf. Biber *et al.*, 1999: 19), their use of aspect is strikingly similar – the curves for the distribution of aspect markers for FLOB and Frown are almost identical to each other (see Fig. 1).

Chinese and English, however, do show some differences in the distribution of aspect markers, as shown in Fig. 2. The figure shows the frequencies of aspect markers, as percentages, in the fifteen text categories in the three corpora. As can be seen, by comparison to the two major varieties of English, aspect markers in Chinese occur more frequently in categories G and K but less frequently in N, L, H, and E. The relatively low frequency of aspect markers in category N (martial arts fiction) in relation to other fiction types, as noted in Section 1, is shown even more markedly in the contrast of the N category between LCMC and FLOB/Frown. British English and American English also differ in that the latter variety does not show such a marked fluctuation in aspect marking in narrative texts, notably in biography and the five types of fiction.

15.4 Kilpiö (1997)

Kilpiö (1997) examines, on the basis of the Helsinki corpus (see unit 7.7), two distinct areas connected to verb BE: developments in its morphology and developments in its functional load from Old English (OE) to Early Modern English (EModE). This excerpt discusses the developments in the functions of BE. The data covers four sub-periods OE1-4 (OE1: -850 A. D., OE2: 850-950 A. D., OE3: 950-1050 A. D., OE4: 1050-1150 A. D.), ME1 (1150-1250 A. D.), ME3 (1350-1420 A. D.) and EModE1 (1500-1570 A. D.).

Kilpiö, M. 1997. 'On the forms and functions of the verb *be* from old to modern English'. In M. Rissanen, M. Kytö and K. Heikkonen (eds.) *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles*, pp. 101-120. Berlin: Mouton de Gruyter.

3.3. Developments in the functions of *be*

In his discussion of the use of *be* as a tense auxiliary, Mustanoja makes the following remark: "It is perhaps not without significance that while *be* is becoming an auxiliary *par excellence* of the passive voice, it is losing ground as an auxiliary of the perfect and pluperfect tenses" (Mustanoja 1960: 501). There is indeed good reason to assume that developments in different parts of the verbal system are not separate and autonomous but interdependent.

3.3.1. Chronological trends in the relative share of the main functions of *be*: a survey of present tense form from OE to EModE1

Table 6 presents an overview of the relative share of the three main uses of the verb *be* from Old to Early Modern English in the present tense, indicative and subjunctive.

The overall impression gained from the statistics in Table 6 is one of great stability in the relative share of the different functions of *be* throughout the periods studied. For a discussion of the implications of this, see section 3.3.

Table 7 gives the breakdown of the auxiliary uses of *be* in the periods studied, throughout which the use of *be* as a passive auxiliary is the most important. With rare constructions like the progressive the method of sampling adopted here clearly involves a random factor. As the corresponding eight OE examples have been classified as copular constructions, only ME3 contains examples of the progressive (see, however, Table 9 below for EModE1 instances of the progressive).

As can be seen, neither the *be to* construction, illustrated above by (10), nor the *be about to* construction are common in any of our periods. The rather high OE percentage is evidently due to the commonness in OE of the (particularly deontic) construction of the type seen in (19):

(19) Nu ge habbað gehyred anrædlice hwæt eow *to donne is* and hwæt eow *to forgane is*. (Ælfric, *Letter to Wulfsige* 34)

'Now you have heard definitely what you are to do and what you are to abstain from.'

The rise of *be to* in EModE1 after being rather dormant in the two ME periods is in accordance with Mustanoja (1960: 524), who says that the "construction is comparatively infrequent in OE and early ME, but becomes more common in later ME and early ModE."

3.3.2. The relative share of the main functions in ME3 and EModE1 in finite past tense forms and in non-finite forms

As stated above on p. 106 in connection with the statistics presented in Table 6, the overall impression gained from a survey of the relative share of the three main functions of *be* is one of great stability. Particularly with regard to the transition from Middle English to Early Modern English this runs counter the expectations that the relative share of auxiliary uses at the expense of the remaining two uses would rise. It is for this reason that I here supplement the information provided by Table 6 by considering the main functions of *be* in past tense forms and in non-finite forms of *be* for the last two subperiods, ME3 and EModE1.

Table 6. Relative share of the three main functions of *be* in OE, ME1, ME3 and EModE1 in the select corpus (every 10th instance): present indicative and subjunctive.

	OE		ME1		ME3		EModE1	
	N	%	N	%	N	%	N	%
Copular uses	527	65%	176	68%	276	64%	235	64%
Auxiliary uses	231	28%	59	23%	119	28%	103	28%
Main verb, non-copular	56	7%	22	9%	34	8%	31	8%
Total instances	814		257		429		369	

Table 7. Types of auxiliary uses.

	OE		ME1		ME3		EModE1	
	N	%	N	%	N	%	N	%
Passive auxiliary	202	88%	52	88%	95	80%	93	90%
Pass. or tense auxiliary	19	8%	1	2%	7	6%	4	4%
Tense auxiliary	10	4%	6	10%	12	10%	6	6%
Progressive (ME-)	–	–	0	0%	5	4%	0	0%
Totals	231		59		119		103	

Table 8. Quasi-auxiliary uses of *be* included in the copular instances.

	OE	ME1	ME3	EModE1
Be to	19	3	3	7
Be about to	–	–	–	1

Percentage of *be to* constructions of all copular constructions:
OE 3.6%, ME1 1.8%, ME3 1.1%, EModE1 3%

Table 9. Relative share of the three main functions of *be* in a select corpus (every 8th instance)[a] of past tense forms of *be* in ME3 and EModE1 and the breakdown of the auxiliary uses between different types of auxiliaries.

Function of *be*	ME3	N	%	EModE1	N	%
Copular	120	54%	126	54%		
Auxiliary	82	37%	91	39%		
Main verb, non-copular	19	9%	16	7%		
Totals	221	100%	233	100%		
Auxiliary uses						
– Passive auxiliary	71		81			
– Passive or tense auxiliary	4		3			
– Tense auxiliary	4		4			
– Progressive auxiliary	3		3			
Total auxiliaries	82		91			

[a] Every 8th, not every 10th instance was analysed here. The solution adopted was purely practical since the structure of the WordCruncher program makes the selection of every 8th example speedier and more mechanical than 10th.

There are immediately obvious features of the relative shares of the main functions of *be* in the light of Tables 6 and 9. The first is that with past tense forms, auxiliary uses are relatively more common and copular uses correspondingly less common than with the present tense forms in subperiods ME3 and EModE1. The difference is of the same magnitude in both subperiods studied as appears from the following juxtaposition of percentages. The figures before the slash (/) give the percentage met in the present, the one after the slash the percentage met in the past tense:

ME3: copular 64% / 54%; auxiliary 28% / 37%; main verb, noncopular 8% / 9%.
EModE1: copular 64% / 54%; auxiliary 28% / 39%; main verb, noncopular 8% / 7%.

There is no obvious explanation for the difference between the present and past tense of *be* with regard to the relative frequency of the main uses of the verb.

The second noticeable thing that emerges from Table 9 is that when we move from ME3 to EModE1 there are no big changes in the relative proportions of the main uses; thus the addition of the past tense to the survey does not change the picture of relative stability gained from a study of the present tense forms.

Table 9 also shows the distribution of past tense forms of *be* between different auxiliary uses. The figures can be compared with those in Table 7; they show a similar kind of breakdown where the use of *be* as a passive auxiliary is preponderant (87% of the ME3 and 89% of the EModE1 auxiliary instances of *was*, *were*, etc. represent the passive auxiliary).

Table 10. Relative share of the main functions of *be* in a select corpus (every 10th instance) of nonfinite forms of *be*.

	ME3	N	%	EModE1	N	%
Infinitive						
Copular		43	40%		42	37%
Auxiliary		57	52%		66	58%
Main verb, non-copular		9	8%		6	5%
		109	100%		114	100%
Past participle						
Copular		7	54%		12	60%
Auxiliary		5	38%		7	35%
Main verb, non-copular		1	8%		1	5%
		13	100%		20	100%
-ing form						
Copular		2	100%		13	68%
Auxiliary		0	0%		6	32%
Main verb, non-copular		0	0%		0	0%
		2	100%		19	100%
All non-finite forms						
Copular		52	42%		67	44%
Auxiliary		62	50%		79	52%
Main verb, non-copular		10	8%		7	4%
Total forms analysed		124	100%		153	100%

The data in Table 10 consists of those non-finite forms of *be*, infinitives, past participles and *-ing* forms in ME3 and EModE1 which are used in verb phrases so as to represent one of the three main functions of the verb *be*. Thus, to give a couple of examples, in (20) the infinitive has the function of a non-copular main verb, in (21) the past participle has the function of a passive auxiliary and in (22) the *-ing* form, a present participle, functions as a copula:

(20) Lat *be* soche falsheed; (*The Cloud of Unknowing* 23)
‘Let such falsehood be’
(21) But [in] this thing hath *ben* discoveryd to the [that] thow seydest that thow wistest not a litel herbyforn (Chaucer, *Boethius’ De Consolatione Philosophiae* 436.C2)
‘But in this thing has been revealed to you what you said that you did not know a little before this time’
(22) y^{e} mylner *beyng* w^{t}yn asked who was ther (*A Hundred Mery Talys* 36)
‘the miller, being within, asked who was there’

Compared with the division of finite forms of *be* between the different functions of the verb set out in Tables 6 and 9 above, the breakdown of nonfinite forms seen in Table 10 again presents a different kind of picture. As only infinitives are represented by a large enough number of instances to enable us to make reliable statistical comparisons between the two periods examined, the main focus will be on this non-finite form. It is worth noticing that with infinitives, both in ME3 and in EModE1, the auxiliary uses are the most common function of the infinitive (52% of the infinitive instances in ME3, 58% in EModE1). In the infinitive there is also a rise in the percentage of auxiliary uses when we move from ME3 to EModE1. By the same token, the relative shares of copular and non-copular main verb uses decrease in EModE1 compared to ME3.

It is interesting to note that the finiteness or non-finiteness of the form of *be* affects its distribution between the copular and auxiliary uses. This is understandable in view of the general tendency to increase three-verb groups in 16th century English and the natural avoidance of non-finite copulas of the type illustrated by example (22).

Of the auxiliary instances, the majority represent the passive auxiliary both in ME3 and EModE. Thus of the 57 infinitives used as auxiliaries in ME3 56 are passive auxiliaries and one is a tense auxiliary; all five past participles with auxiliary function represent the passive auxiliary. In subperiod EModE1, all the non-finite forms with auxiliary function represent the passive auxiliary. This confirms the picture gained from finite forms of *be* functioning as an auxiliary.

One feature in Table 10 that points the way to future developments is the great increase in the number of occurrences of *being* used either as a gerund or as a present participle when we move from subperiod ME3 to EModE3. This expansion naturally paves the way to the enrichment of the morphology of the progressive.

15.5 Mair, Hundt, Leech and Smith (2002)

While Kilpiö (1997) traces language in transition over several centuries, Mair, Hundt, Leech and Smith (2002) explore language change occurring over a shorter span of time. This paper compares part-of-speech tag frequencies in two matching one-million-word reference corpora of standard British English, LOB and FLOB (see unit 7.4). The study shows a significant rise in the frequency of nouns, which is not paralleled by a corresponding decrease in verbs. This excerpt examines frequency changes among subcategories and combinations of nouns and provides an explanation from both a diachronic and a synchronic perspective.

Mair, C., Hundt, M., Leech, G. and Smith, N. 2002. 'Short term diachronic shifts in part-of-speech frequencies'. *International Journal of Corpus Linguistics* 7/2: 245-264.

3. Frequency changes among subcategories and combinations of nouns

Leaving aside discussion of other word classes, we may at this stage look more closely at the noun category from yet a further viewpoint: let us consider the frequency of different subcategories of nouns, to find out if the noun increase between LOB and F-LOB is concentrated in one subcategory rather than another.

The striking feature of Table 3, as of Tables 1 and 2, is the consistency of the increase in the use of nouns across different categories and subcategories. However, although all three of these important subclasses of nouns show the same increase, they do so to markedly different degrees. The most significant increase of all is that of proper nouns, which amounts to 11%. Why the texts of F-LOB contain so many more proper nouns than the texts of LOB is not one of the questions to be answered in this article, but one suggestion which may contribute to the answer is that F-LOB reflects a greater prevalence of acronyms in the 1990s, as shown in Table 4.

Most proper nouns which are printed entirely in capitals are acronyms: words such as UNO, UNICEF, RSPCA, etc. Although these do not make up a large proportion of all proper nouns, it is worth noting a remarkable difference between their incidence in the two corpora: acronyms appear to be nearly twice as frequent in F-LOB as in LOB.

We now illustrate another way of attacking the issue of the higher frequency of nouns in F-LOB. This is to obtain counts of noun + noun sequences, to see what change if any has taken place between LOB and F-LOB. There is more than a suspicion that the favoured Germanic way of forming complex lexical expressions – the combining of nouns – is making a comeback in the later 20th century, and it may be further suspected that this change is more salient in newswriting (Press) than in other categories: witness the well-known multiple-noun headlines such as:

BT strike threat over plans to chop 1,000 (F-LOB text A06)

Flagship hospital boss out (F-LOB text A07)

To investigate this, our first tactic was to count all tags N* N*: that is, any noun (including proper nouns) followed by other noun. The results showed a vastly significant increase in the use of noun + noun sequences in F-LOB, as shown in Table 5.

Table 3. Frequency of selected noun subcategories in the LOB and F-LOB corpora

	LOB corpus		F-LOB corpus		Difference	
Subcorpus	Raw freq.	per million	raw freq.	per million	% of LOB	log likelihood
Singular common nouns						
Press	28047	157754	28772	161386	+2.3%	7.4
Gen. Prose	65631	158274	67996	164335	+3.8%	47.2
Learned	27254	169473	27592	172093	+1.5%	3.2
Fiction	32764	127726	34278	133450	+4.5%	32.2
Total	153696	152206	158638	157186	+3.3%	80.9
Plural common nouns						
Press	9214	51825	9835	55166	+6.4%	18.6
Gen. Prose	23844	57501	26117	63119	+9.8%	108.4
Learned	9806	60977	10783	67256	+10.3%	49.4
Fiction	8037	31331	9213	35868	+14.5%	78.7
Total	50901	50407	55948	55436	+10.0%	241.3
Proper nouns						
Press	12246	68879	12413	69626	+1.1%	0.7
Gen. Prose	14432	34804	17579	42486	+22.1%	316.9
Learned	3765	23412	4551	28383	+21.2%	76.7
Fiction	9229	35978	9474	36885	+2.5%	2.9
Total	39672	39287	44017	43614	+11.0%	228.1

Table 4. Proper nouns consisting entirely of capital letters: comparison of frequency in LOB and F-LOB

	LOB corpus		F-LOB corpus		Difference	
Subcorpus	Raw freq.	per million	raw freq.	per million	% of LOB	log likelihood
Press	775	4372	857	4811	+10.0%	3.7
Gen. Prose	391	946	1196	2895	+205.9%	428.1
Learned	98	617	615	3852	+524.1%	414.7
Fiction	166	648	188	731	+12.8%	1.3
Total	1430	1422	2856	2833	+99.2%	479.7

Table 5. Noun + noun sequences: comparison of frequency in the LOB and F-LOB corpora

	LOB corpus		F-LOB corpus		Difference	
Subcorpus	raw freq.	per million	raw freq.	per million	% of LOB	log likelihood
Press	9876	55714	10874	61045	+9.6%	43.3
Gen. Prose	12938	31306	16229	39277	+25.5%	372.8
Learned	5260	33127	5961	37336	+12.7%	40.0
Fiction	4127	16121	4952	19261	+19.5%	71.6
Total	32201	32030	38016	37711	+17.7%	466.3

Table 6. Sequences of Noun + Common noun: comparison of the LOB and F-LOB corpora (excluding tags NNB, NNL*, and NNA, which are invariably associated with naming expressions)

	LOB corpus		F-LOB corpus		Difference	
Subcorpus	Raw freq.	per million	raw freq.	per million	% of LOB	log likelihood
Press	5098	28760	6376	35794	+24.5%	136.5
Gen. Prose	8756	21187	11562	27982	+32.1%	389.4
Learned	4459	28083	5235	32788	+16.8%	58.0
Fiction	2448	9562	3366	13092	+36.9%	141.7
Total	20761	20651	26539	26326	+27.5%	691.9

Strikingly, the most dramatic increases of noun + noun sequences are not found in Press (A-C), where it could be expected, but rather in other categories, particularly General Prose. It was decided to try other variants, but surprisingly, it was not combinations ending with a proper name, but combinations ending with a common noun that showed the steepest increase of occurrence. In Table 6, we compare LOB and F-LOB in terms of sequences of noun + common noun.

The table shows a very marked difference – an increase of 27.5% in F-LOB above the frequency in LOB. Note that the Noun + Common noun rise is a feature of every text category A-R, not just the four block groupings used in this paper; whereas Noun + Proper Noun sequences rise in only 6 of the 15 text categories.

4. Shifts in part-of-speech frequencies: Diachronic and synchronic factors

To cast further light on tag frequency in a diachronic perspective, it is instructive to relate the observed changes to the synchronic variation manifest in a given corpus at any one time. In their exhaustive analysis of the tagged LOB corpus, Johansson and Hofland, for example, have shown tag frequencies to vary quite drastically from genre to genre (1989/I:7–39, in particular 15). Our figures, which are based on the C8 re-tagging of LOB and therefore differ from theirs in minor ways, are as follows:

Table 7. Noun and verb frequencies in LOB (given as percentages)

	Nouns	verbs
Fiction	20.0	21.9
Nonfiction (all)	26.9	16.4
Nonfiction / press (A-C)	29.6	16.6
Nonfiction / science (J)	26.2	15.5
Total	25.1	17.8

In the wake of Johansson and Hofland’s pioneering effort there have been a number of further corpus-based studies of part-of-speech distribution – most recently Biber et al.’s (1999) *Longman Grammar of Spoken and Written English*. None of them – including Hudson’s (1994) facetiously titled “About 37% of Word-Tokens are Nouns” – casts doubt on the strong tie between genre/text-type and the frequency of nouns and verbs.

Stated in the most simple terms, the major result of all such research is the following: information orientation appears to promote the use of nouns, whereas narration is characterised by a higher incidence of verbs. LOB does not contain any spoken language, so that it is impossible to ascertain without further data analysis to what extent the results from the Fiction (K-R) sections, through the incorporation of fictional dialogue, represent the situation in speech. However, Leech et al. (2001: 294–295) gives comparative percentages for the frequency of nouns and verbs as in Table 8, demonstrating that the high verb-to-noun ration shown for fiction in Table 7 is even higher in general spoken corpus material.

What does all this mean in terms of the diachronic analysis attempted in the present paper? First and foremost, the extent of the synchronic variation observed makes clear that smallish shifts in part-of-speech ratios over time must be interpreted with extreme caution. After all, what is the significance of a 5.3% increase in nouns in the corpus overall, when at any given time there is a much greater scope for variation based on genre?

Table 8. Noun and verb frequencies in the BNC sampler (given as percentages)

	Nouns	verbs
Written texts	28.4	17.3
Spoken transcriptions	14.6	23.1

Changes in tag frequencies thus do not reflect grammatical change directly. Rather, they may hold a clue to the puzzle of how grammatical innovations spread in actual usage, namely at differential speeds through different genres. To illustrate this general assumption, consider a concrete case at hand, namely the rise in verbs of 7.3 per cent observed in our reportage samples (sections A in LOB and F-LOB). This is not a direct sign of a grammatical change, but shows a style change. Reportage over the past thirty years has moved a little closer towards other genres rich in verbs – represented by fiction and conversation in our corpora. Such colloquialisation and informalisation of news writing is a sociocultural rather than a linguistic phenomenon – and has been plausibly accounted for by critical discourse analysts, sociologists and historians (cf., e.g., Fairclough 1992). But in due course, it will no doubt have consequences for the linguistic system, because the new stylistic climate will speed up the demise of many lexical and grammatical archaisms and prevent the establishment of new lexical and grammatical markers of more formal or literary diction.

Standard English is primarily defined through its lexicon, and through its grammar. On a textual level, however, standard English is also usage, style and choice. This is, after all, the level on which we immediately recognise the standard British English of the beginning of the 20th century and distinguish it from 1960s and 1990s English, or tell British standard English apart from American standard English – long before we confirm such first intuitions through laborious counts of grammatical or lexicogrammatical variables such as the proportion of analytical and synthetic comparatives/superlatives or the prevalence of regularised *spoiled* and *burned* against their irregular counter-parts *spoilt* and *burnt*. At this level of language change – for lack of a better term one might speak of changes in grammar-in-text – the comparison of tag frequencies will usefully complement the quantitative study of lexical frequencies and the qualitative analysis of individual examples. In addition, the study of changing stylistic fashions and genre conventions is an interdisciplinary undertaking, linking linguistics, sociology and cultural history. The investigation of corpora may thus yield insights which are useful far beyond the field of linguistics itself, and this is a prospect we need not be unhappy about at all.

15.6 Unit summary and looking ahead

This unit demonstrated the use of corpora in contrastive and diachronic studies, with particular reference to multilingual corpora and diachronic corpora. Readers are reminded that while parallel corpora are useful in translation studies, they are typically complemented by comparable corpora when used in contrastive studies (see unit 10.6). As we will see in case study 6 in Section C, translated language is distinct from L1 language. It was also noted that diachronic studies would not really have been possible without corpus data. Diachronic corpora are useful in tracking developments in the syntactic, semantic and functional distributions of linguistic features in both the long and short terms. In case study 2 readers will have an opportunity to explore how language change over three decades (from the early 1960s to the early 1990s) has influenced speakers' choice between a *to*-infinitive and a bare infinitive following *HELP*. In the next unit, the final unit in Section B, we will demonstrate the use of corpora in an important area of linguistics – language teaching and learning.