# COMPUTER CORPORA AND THEIR USE IN LANGUAGE ANALYSIS

Atsuko Furuta UMESAKI

## 1. INTRODUCTION

Recent development of computer technology has enabled us to obtain a massive amount of linguistic data from texts of language in actual use. A collection of these texts in a machine-readable form is called a (computer) corpus. Up to now, most linguistic theories proposed have normally been based on qualitative data propounded by researchers, the representativeness of which depends on native speakers' intuition. It can be expected that the use of corpora will make quantitative as well as qualitative analysis. The present study gives an introduction to existing corpora, and consideration is made of the possibilities and limitations of their use in linguistic research. A personally-made corpus is also dealt with in the present article. An analysis of nominalised forms made by the present writer is given as an illustration of the possibilities now available.

## 2. CORPUS-BASED STUDIES

According to Aijmer and Altenberg (1991, p. 1) the study of language on the basis of text corpora is called 'corpus linguistics'. Akano (1989, p. 142) gives a more concrete definition of 'corpus linguistics' by saying that it consists of two major domains : the study of techniques and methods concerning the corpus construction and that of language by the use of corpora.

In terms of the study of language, Leech (1992, p. 107) points out four features of corpus linguistics : (1) its focus on linguistic performance rather than competence, (2) its focus on linguistic description rather than linguistic universals, (3) its focus on quantitative as well as qualitative models of language, (4) its emphasis on an empiricist rather than rationalist view of scientific inquiry.[1]

### 2.1 Corpora

A corpus is defined by Crystal (1987, p. 410) as a representative sample of language compiled for the purpose of linguistic analysis. Akano et al. (1991, p. 1) define it as a collection of spoken and/or written texts representing a certain language, a dialect or other variety, treated and stored in machine-readable form to serve the purpose of linguistic study.[2] According to Saito (1992, p. 5) it is a body of texts compiled in machine-readable form for the purpose of linguistic and literary research.

The basic definition of corpus is a collection of texts, but with recent development of computers, this has come to entail machine-readable as part of the definition. Another notable feature of the cor-

**Table 1** Kinds of Corpora

<table>
<tr><td>Descriptive Comments</td><td colspan="9">Kinds of Corpora</td></tr>
<tr><td>broader understanding</td><td colspan="9">Corpus of actually spoken/written language in machine-readable form</td></tr>
<tr><td>narrower understanding</td><td colspan="8">Corpus of samples representing varieties of language use designed for linguistic research</td><td>Corpus of a fixed size without sampling representative texts</td></tr>
<tr><td>closed/open-ended nature</td><td colspan="7">sample corpus</td><td>monitor corpus</td><td></td></tr>
<tr><td>period covered for texts collected</td><td colspan="6">synchronic</td><td>diachronic</td><td></td><td></td></tr>
<tr><td>medium of language</td><td colspan="6">written and/or spoken</td><td>(written)[1]</td><td></td><td>(written)[2]</td></tr>
<tr><td rowspan="2">purpose of sampling</td><td colspan="2">general purpose</td><td colspan="4">special purpose</td><td rowspan="2">multi-genre, one-genre/ one-author, regional corpora[3]</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td colspan="2">multi-genre corpus of a regional variety</td><td colspan="2">domain-specific corpus</td><td colspan="2">corpus of a social / non-standard regional dialect</td></tr>
<tr><td>Examples of existing corpora</td><td colspan="2">Brown, LOB, LLC, Survey of English Usage Corpus</td><td colspan="2">Guangzhou Petroleum English Corpus</td><td colspan="2">PoW Corpus (children's speech)</td><td>Helsinki Corpus</td><td>—[4]</td><td>CD-ROM of <i>the Guardian</i></td></tr>
<tr><td colspan="10"></td></tr>
<tr><td rowspan="2">International Corpus of English</td><td rowspan="2">core corpus</td><td colspan="5">optional corpus</td><td rowspan="2">—[6]</td><td rowspan="2">monitor corpus</td><td rowspan="2"></td></tr>
<tr><td colspan="2">expanded corpus</td><td colspan="2">specialised corpus</td><td>corpus of scl./non-stnd. dialects[5]</td></tr>
<tr><td>COBUILD Project</td><td colspan="5">main corpus<br>reserve corpus</td><td>—[7]</td><td>—[8]</td><td>monitor corpus</td><td></td></tr>
</table>

[1] Since recorded speech has existed for only half a century, the compilation of a diachronic spoken corpus may be impractical.

[2] If texts are transcribed, the corpus may be regarded as domain-specific since transcription is normally carried out for the purpose of linguistic analysis.

[3] Diachronic corpora are categorised by Kytö et al, eds. (1994).

[4] There exist no independent examples of a monitor corpus.

[5] 'scl./non-stnd. dialects' stands for social/non-standard dialects. Greenbaum (1992, pp.172–173) describes this category as 'collection of speech to illustrate nonstandard sociolects or regional dialects, the language of children, or the language of immigrant communities'. However, as these languages may not be restricted to speech, Greenbaum's terminology is not adopted here.

[6] Neither Greenbaum (1992) nor Leitner (1992) touch upon a diachronic corpus in the concept of ICE although it may be included in optional corpora.

[7, 8] Not mentioned, but may possibly be developed.

pus is that it is 'essentially a body of natural language material' (Leech et al., 1992, p. 115) so that it provides authentic data for linguistic research. It can therefore be said that a corpus is a collection of actually spoken/written language compiled in machine-readable form.

A number of corpora have been compiled in various parts of the world by assembling representative samples for linguistic research. Thirty-six English corpora are listed in Johansson and Stenström, eds. (1991, pp. 319–354), and seventeen in Aijmer and Altenberg, eds. (1991, pp. 315–318). Apart from these corpora, however, there are compilations of literary works and newspapers in machine-readable form, some of which are distributed commercially.[3] These are normally not representative selections but complete works, not specially designed for linguistic research. However, they can be made use of for linguistic study, and therefore are regarded as corpora in the present study.[4]

Corpora can be classified to give a clearer picture of their basic characteristics. In terms of size, two kinds of corpora can be envisaged : a sample corpus[5] and a monitor corpus. A sample corpus is of fixed size and usually contains relatively short samples, while a monitor corpus is open-ended and consists of complete texts (Clear, 1992, p. 28). The monitor corpus is gigantic, an ever-changing store of texts for the purpose of making routine records (Sinclair, 1991, p. 25).

Sample corpora can be subdivided in terms of the period covered for the texts collected, that is, whether the corpus is synchronic or diachronic. In addition, the former can be made up of written and/or spoken texts, while the latter will almost certainly consist of written texts only, since recorded speech has existed only for the last fifty years.

Synchronic corpora have either a general or a special purpose.[6] A general purpose corpus consists of texts representing a number of registers, and can be regarded as representative of a regional variety. Special corpora may be, for example, domain-specific[7] or corpora of a social dialect or non-standard regional dialect.[8] Diachronic corpora are categorised as 'multi-genre corpora', 'one-genre or one-author corpora' and 'regional corpora' in Kytö et al. (1994, p. vii).

Table 1 attempts to summarise the above categorisation, giving examples of existing corpora, suggesting possibilities along the lines of present developments. Although there are no good examples of monitor corpora, as is pointed out by Clear (1992, p. 28), part of the work of the COBUILD project can be seen as a monitor corpus. The Brown, LOB and London-Lund corpora can be regarded as typical examples of the sample corpus and of corpora of regional varieties with a general purpose. These three corpora are contained in a CD-ROM, *ICAME Collection of English Language Corpora*[9] together with two other corpora.

### 2.1.1 The Brown Corpus (The Standard Corpus of Present-Day Edited American English)

This corpus was compiled at Brown University from 1963 to 1964. It contains 500 written texts, each consisting of about 2000 words, approximately one million words in total. The 500 texts were taken from books, journals, magazines and newspapers published in the U.S.A. in 1961, representing 15 categories or genres.[10]

### 2.1.2 The LOB Corpus (The Lancaster–Oslo/Bergen Corpus of British English)

Regarded as a British English equivalent of the Brown Corpus, the LOB Corpus was compiled at the University of Lancaster and the University of Oslo from 1970 to 1978. It contains 500 written texts, each consisting of about 2000 words, approximately one million words in total. The texts were taken from books, journals, magazines, newspapers published in Great Britain in 1961, representing 15 categories. The number of texts in each category of the LOB Corpus is almost the same as for the Brown Corpus.

**Table 2** Structure of the Brown and LOB Corpora[11]

| Text Category (Genre) | | | Number of | |
|---|---|---|---|---|
| | | | Texts | Words(ca.) |
| Informative prose | A | Press : reportage | 44 | 88,000 |
| | B | Press : editorial | 27 | 54,000 |
| | C | Press : reviews | 17 | 34,000 |
| | D | Religion | 17 | 34,000 |
| | E | Skills, trades and hobbies | 38* | 76,000 |
| | F | Popular lore | 44* | 88,000 |
| | G | Belles lettres, biography, essays | 77* | 154,000 |
| | H | Miscellaneous<br>(government documents, industry reports, etc.) | 30 | 60,000 |
| | J | Learned and scientific writings | 80 | 160,000 |
| Imaginative prose | K | General fiction | 29 | 58,000 |
| | L | Mystery and detective fiction | 24 | 48,000 |
| | M | Science fiction | 6 | 12,000 |
| | N | Adventure and western fiction | 29 | 58,000 |
| | P | Romance and love story | 29 | 58,000 |
| | R | Humour | 9 | 18,000 |
| | | TOTAL | 500 | 1,000,000 |

*The number of texts in categories E, F and G of the Brown Corpus are 36, 48 and 75, respectively. (Table 2 in Oostdijk, 1991, p. 37)

The LOB Corpus have untagged and tagged versions, available through the ICAME.[12] Grammatical tagging was undertaken by the automatic tagging system called CLAWS and by manual pre- and post-editing (Garside et al., eds. 1987, Johansson et al. 1986). The following provides an example.

untagged version

A 01 16 |^He believes that the House of Lords should be abolished and that

A 01 17 Labour should not take any steps which would appear to* "prop up**" an

A 01 18 out-dated institution.

tagged version

A 01 16 ^ he_PP 3A believes_VBZ that_CS the_ATI House_NPL of_IN Lords_NPTS

A 01 16 should_MD be_BE abolished_VBN and_CC that_CS

A 01 17 labour_NN should_MD not_XNOT take_VB any_DTI steps_NNS which_WDTR

### 2.1.3 The London–Lund Corpus of Spoken English

This corpus contained 87 Spoken British English texts, each consisting of about 5000 words when it was completed at Lund University in 1979. Later 13 texts were added and now the corpus consists of 100 texts, about 500,000 words in total. The texts were recorded from 1953 to 1988. The composition of the corpus is shown in Table 3.

**Table 3** Composition of the London-Lund Corpus

| Biber's Classification | Number of Texts (5000 words/text) | | | Classifications in London-Lund Corpus |
|---|---|---|---|---|
| Face to face conversations | S.1 | 14 | 14 | conversations between intimates and distants, surreptitiously recorded |
| | **S.2 | 14 | 14 | 〃 |
| Face to face conversations | S.3 | 6 | 7 | 〃 |
| | **S.4 | 7 | 7 | conversations between intimates and equals, wholly non-surreptitious or composite |
| Interviews* | *S.5(1-7) | 7 | 7 | public conversations between equals, non-surreptitious |
| | (8-11) | 4 | 6 | private conversations between equals, non-surreptitious |
| Interviews* | *S.6 | 6 | 9 | conversations between disparates, non–surreptitious |
| Telephone conversations | S.7 | 3 | 3 | telephone conversations between personal friends, surreptitious |
| 〃 | S.8 | 4 | 4 | telephone conversations between business associates, surreptitious |
| 〃 | S.9 | 3 | 5 | telephone conversations between disparates, surreptitious |
| Broadcasts | S.10 | 8 | 11 | spontaneous commentary |
| Spontaneous speeches | S.11 | 5 | 6 | spontaneous oration |
| Prepared speeches | S.12 | 6 | 7 | prepared but unscripted oration |
| | Total | 87 | 100 | |

*Biber (1988, pp. 209–210, 264–269) classifies S.5 (1–7) and S.6 (excluding S.6 (2) and S.6 (4 b)) into one group, 'Public conversations, debates, interviews,' represented by the term 'Interviews'.
**S.2 and S.4 are not used for his analysis.

Prosodic features are assigned to the texts of the London-Lund Corpus. Tone units, nuclei (indicating the direction of pitch), boosters (indicating the range of pitch), stresses and pauses are shown by symbols, although other features such as tempo, loudness, voice quality are omitted (Svartvik et al. 1982, p.19). The effects the symbols[13)] indicate can be realised by a comparison of the text in the printed book (ibid, pp. 21–28) and the equivalent part stored in the computer being used. For example, in the Macintosh format illustrated below, # indicates a tone unit boundary ; ^ an onset ; ¥ a falling tone, / rising tone, a combination of ¥/ a fall-rise tone, = a level tone in nucleus ; – and –– pauses ; ! a booster indicating higher than preceding pitch-prominent syllable ; : a booster indicating higher than preceding syllable ; " heavy stress ; **simultaneous talk ; ( ) contextual comment ; @ a schwa as part of a hesitation marker.

```
1 3 9 1510 1 1 A 11 and em^br/oiders# /
1 3 9 1520 1 1 A 11 and *^d/arns# /
1 3 9 1530 1 1 A 11 and sews* ^b¥uttons on# /
1 3 9 1540 1 1 b 20 *(–laughs) yes* /
1 3 9 1550 1 1(A 11 ––and I ^s=aid# /
1 3 9 1560 1 1 A 11 well I ^don`t r¥eally _think# /
1 3 9 1570 1 1 A 11 I could ^wr¥ite# –– /
1 3 9 1580 1 1 A 11 and this was a sort of ^ninety–six page  : b¥ooklet# /
1 3 9 1590 1 1 A 11 ^you kn/ow# /
1 3 9 1600 1 1 A 11 about ^that b¥ig# *–* /
1 3 9 1610 1 1 A 11 [@m] I`d I`d ^need to g¥o through# /
1 3 9 1620 2 1 A 21 ^each of the /

1 3 10 1640 2 1 A 21 I don`t think it will be e^nough just to have/
1 3 10 1650 1 1 b 20 *[m]* /
1 3 10 1640 1 1(A 11 them !d¥/emonstrated#. /
1 3 10 1660 1 1 A 11 and then ^I`ll !wr¥ite it you see#. /
1 3 10 1670 1 1 A 11 so they “^sent – the machine ov¥er# /
1 3 10 1680 1 1 c 20 (enters)
```

### 2.1.4 Other Sample Corpora

Besides the Brown, LOB and London-Lund corpora, the following two corpora are stored in the CD-ROM (see 2.1)[14]. One is the Kolhapur Corpus, compiled at Shivaji University from 1980 to 1986. This is a written Indian English corpus consisting of approximately one million words. 500 texts, taken from materials printed in 1978, are divided into the 15 categories with about 2000 words per text. The categories of the Kolhapur Corpus are the same as those of the Brown Corpus although the number of texts and sub-categories are different.

The other is the Helsinki Corpus of English Texts (Diachronic Part) consisting of about 1.6 million words in total, samples of text dating from the 8th to the beginning of the 18th century (OE : 413,250 words, ME : 608,570 words, EModE : 551,000 words, according to Saito (1994)). The structure of the corpus is given in Table 4. This corpus was compiled at the University of Helsinki from 1984 to 1991.

### 2.1.5 Corpora of Massive Scale

As Saito (1994 pp. 546–548) remarks, a recent trend is the compilation of massive-scale corpora.

One of these is a corpus collecting varieties of English from all over the world. The International Corpus of English began to be constructed on the initiative of University College London in

**Table 4** Text prototypes and text types in the diachronic part of the Helsinki Corpus of English Texts. (Kyto et al. eds., 1988, p. 175)

| Text Prototype | Text Type | | |
|---|---|---|---|
| | Old English | Middle English | Early Modern English |
| 1. Stipulation | Law | Law | Law |
| | | | Local |
| | | | National |
| | Document | Document | |
| 2. Science | Astronomy | Astronomy | Astronomy |
| | | Medicine | Medicine |
| | | Science Other | Science Other |
| 3. Instruction | Recipe | | |
| Secular | Prognostication | | |
| | | Handbook | Handbook |
| | | | Education |
| 4. Instruction | Homily | Homily | |
| Religion | | Sermon | Sermon |
| | Rule | Rule | |
| | Religion. Treatise | Religion. Treatise | |
| 5. Narration | History | History | History |
| Non-Fictive | | | Travelogue |
| | | | Diary |
| | | | Private |
| | | | Non-Private |
| | | | Biography |
| | | | Auto-Private |
| | | | Other |
| 6. Narration | Fiction | Fiction | Fiction |
| Fictive | | Romance | |
| | | Biography | |
| | | Life Saint | |
| | | Travelogue | Travelogue |
| 7. Correspondence | Preface | Letter | Letter |
| | | Private | Private |
| | | Non-Private | Non-Private |
| 8. Drama | | Play | Play |
| | | Miracle | Comedy |
| | | Morality | |
| 9. Proceeding | | | Trial |
| | | | Meeting |
| | Deposition | Deposition | Deposition |

1990. For the purpose of comparing regional varieties, it contains at least fifteen one-million-word corpora representing regional varieties, and consists of spoken and written texts with identical text categories. The concept of a core corpus is similar to that of a general purpose corpus. According to

Greenbaum (1992, pp. 172–173) the International Corpus of English has four optional corpora besides a core corpus : an expanded corpus (where the same categories are retained as in the core corpus but the corpus is enlarged in size) ; a specialized corpus (constituting a particular category which does not exist in the core corpus) ; a corpus of non-standard varieties (such as the language of immigrant communities) ; and a monitor corpus.

The British National Corpus is also a large-scale corpus, consisting of only British English, 100 million words in total (ninety million words of written texts and ten million words of spoken texts). This three-year project started in 1991 in collaboration with Oxford University Press, the Longman Group, Chambers, Lancaster University, Oxford University Computing Services, and the British Library. Texts of 'informative prose' are taken from books, periodicals and so on, printed after 1975 ; and 'imaginative prose' consists of 25% of texts printed from 1960 to 1974 and 75% of those printed after 1975. Half of the spoken texts are made up of everyday conversation demographically sampled at 36 points across Britain and recorded for about 1500 hours in total. Word-class tags are assigned to the texts. (Nakamura (1994), Saito (1994), Leech et al. (1994), Akano (1995))

Another gigantic corpus is the Bank of English, consisting of over 200 million words of mostly British English, 25% of it American English and 5% other native varieties. This corpus is made up of texts from books, newspapers, magazines, brochures, leaflets, reports, letters, radio broadcasting and natural conversations. They constitute several sub-corpora, such as Book Corpus, *Times* Corpus, BBC Corpus, NPR (American National Public Radio) Corpus and Spoken Corpus (consisting of face-to-face conversation, telephone conversation, broadcast interviews and discussions etc. of about 3.6 million words). The Book Corpus has 318 texts extracted from books published from 1983 to 1992, and the average size of texts is 60,000 words. Newspapers represented are *The Independent, Today* and *The Sun* as well as *The Times*. The Bank of English is an outcome of the COBUILD (Collins Birmingham University International Language Database) Project started in 1980, from which the Birmingham Corpus (in 1985) and some corpus-based dictionaries derived. (Inoue (1994), Nakamura (1994), Saito (1994), Akano (1995)) Twenty million words of the texts in the Bank of English are accessible via the Internet at present.[15]

**2.2 Ways of Using Corpora**

As has been said, each corpus has its own characteristics. This enables linguists to analyse language not only in terms of lexis (such as the meaning of words, collocation, morphology and philology) and grammar (e.g. tense and aspects, complexity of phrases), but also from other various linguistic approaches, such as discourse structure (e.g. theme/rheme organisation), register (e.g. spoken and written differences), regional varieties (e.g. American English, British English), social varieties (e.g. children's language) and the historical development of language. Various kinds of research have been conducted, and publications of studies based on corpora[16] have been increasing in number since the making of the first sample corpus-the Brown Corpus. Corpora and these corpus-based studies have been utilised for the compilation of dictionaries as described in 2.1.5. They can also be used for computer-assisted language education.[17] In addition, corpora may be useful not only for linguistic research but also for literary analysis.

### 2.3 Variety, Register, Genre and Text Type

It is generally understood that there are two types of language varieties, namely varieties according to use and according to users.[18] The former can be considered registers or diatypic varieties and the latter dialects or dialectal varieties (Halliday and Hasan, 1985, p. 43). Dialectal varieties can be sub-categorised as regional and social varieties.

'Register' is defined by Halliday and Hasan (1985) in terms of 'field', 'tenor' and 'mode'. The field of discourse concerns subject-matter, the tenor of discourse refers to the relations among participants in the discourse, and the mode of discourse applies to the medium : the choice of speech and writing is included in 'mode'.

'Genre' is defined in various ways.[19] According to Biber (1988, p. 70) it refers to text categorisations assigned on the basis of external criteria, in contrast with 'text type' which refers to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories. In his example, a science fiction text represents a genre of fiction (relating to author's purpose), but it might also represent an abstract and technical text type (in terms of its linguistic form). In Biber's definition, 'genre' is considered to be interchangeable with 'register'.

In applied linguistics, the terms 'genre analysis' and 'register analysis' are used differently. Swales (1981 a, p. 10) claims that traditional register or sub-register analysis differs from 'genre analysis' in the importance attached to communicative purposes within a communicative setting. Hopkins et al. (1988) give the example of discussion in scientific articles, which is a genre, whereas academic prose is a category of register. However, from this definition, it can be understood that 'genre' is simply more specific than 'register', and the definition of 'register' can cover that of 'genre'. Widdowson (1983, p. 101–102) points out that it is not entirely clear just what the term 'genre' is meant to cover.

'Text type' and 'text category' are sometimes used interchangeably in the presentation of corpus structure. The Helsinki Corpus uses the terms 'text proto-type' and 'text type', which are considered to be equivalent to the term 'text category' in the LOB and Brown Corpora. (See Tables 2 and 4)

In the present study, the term 'genre' is understood as text categorisation based not on the similarity of linguistic features but on external criteria and can be defined by the three criteria of 'register'. Genres sometimes accord with text categories in a corpus, and sometimes with a combination or a part of the text categories. 'Text type' is used as text categorisation based on the similarity of linguistic features except when the existing corpus uses the term to refer to text categories.

### 2.4 Theoretical Framework and Corpus Structure

What should be kept in mind in the study of varieties is that categorisations and internal structures of corpora do not necessarily accord with the above theoretical concepts. With regard to regional varieties, both standard and non-standard varieties are construed as dialects with equal linguistic value, but standard and non-standard regional varieties are separately categorised in the corpus compilation, especially in the International Corpus of English. (See Table 1) In the case of register, sample texts of the Brown, LOB and London-Lund Corpora are classified as genres on an intuitive basis, as Oostdijk (1991, p. 40) points out. Therefore, in register analysis, it should always be taken into account whether groups of texts are worth comparing to examine difference of registers.

### 2.5 Ways of Analysing Corpora

Basically there are two approaches to corpora. One is qualitative analysis, in which samples of a target form are extracted from the corpora and each sample is examined. The other is quantitative analysis, for which frequencies of a target form, and related forms in some cases, are calculated. For both kinds of analysis, computers are useful for finding target forms. Searching, frequency calculation and data storing can be carried out in various ways. For these tasks an increasing number of package programs have become commercially available recently. There are also programs which can be downloaded via e-mail.

One of the ways of fulfilling those tasks is to make a KWIC (Key Word In Context) concordance, a word frequency list, a word index and so on.[20] Another way is to extract and store sentences containing a target form.[21] This is equivalent to collecting sample sentences manually on paper cards. The third way is to use Editor programs.[22] For counting occurrences and processing other data, a script processing program such as Sed can be used, and it is also useful to write a small program in simplified computer language such as JGAWK.[23] It is also possible to combine some of these methods.

Statistical analysis will be an effective tool for understanding quantitative data.[24] Recently a number of package programs have become available.[25] Statistics functions effectively in recognising text typology, as shown in the research carried out by Nakamura (1993).

### 2.6 Personal Corpus Compilation

Since researchers have their own aims and purposes, the most suitable material for analysis is a corpus made by themselves. Thanks to the development[26] of OCRs (Optical Character Reader) and scanners, printed scripts can be made machine-readable more easily than before. The existing general-purpose corpora can be used to recognise the characteristics of texts in the personally made corpus by comparing occurrences of some linguistic items in the existing corpora and in the personally made corpus. For the comparison, Biber's data (1988) are useful.

## 3. LANGUAGE ANALYSIS

It is, in general, difficult to pinpoint the factors which determine the selection of one form from the possibilities available. However, factors affecting a syntactic choice can be considered from the following four linguistic viewpoints : morpho-syntax, discourse, dialectal varieties and historical development of language. Factors at the morpho-syntactic level, including phonological factors, lie in differences of meaning possible for the same sentence. Factors at the discoursal level are mostly related to register, text structure and rhetoric. Dialectal varieties consist of regional varieties, for instance American English or British English, and social varieties such as sex differences. A form may be predominant for a certain period of time, but its use may decline after that. Therefore the use of a form can also be affected by the period when it is used.

A corpus-based study will enable us to make an extensive and synthetic approach from the four factor levels (see 2.2). In the subsequent sections examples of analyses will be given concerning the selection of one nominalised forms instead of other possibilities.

## 4. ANALYSIS OF NOMINALISATIONS

### 4.1 Transitive Verb Nominalisations

The use of the following three transitive verb nominalisations was examined :

(a) *-tion/-ment/*etc.* + *of* + O** (e.g. *the construction of a new building*)

*derivative nouns except for nouns with an *-ing* suffix.

**'O' refers to a phrase functioning as 'object' in the equivalent verbal construction.

(b) *-ing* + *of* + O (e.g. *the constructing of a new building*)

(c) *-ing* + O (e.g. *constructing a new building*)

The target forms were extracted according to the following procedure from Categories J (Academic Prose), A (Press Reportage) and K (General Fiction) of the LOB Corpus and S. 1 and S. 2 (Face-to-face Conversation) of the London-Lund Corpus. A *VZ Editor* program was used for searching the forms and for storing data. With regard to form (a), '*of*' was searched, and then only forms with *-tion/-ment/* etc. derivative nouns preceding *of* were identified manually, excluding the case in which the derivative noun depicts a concrete object and in which a noun phrase after *of* functions as semantic subject of the derivative noun. Similar procedure is taken to extract forms (b) and (c). There are some marginal cases. One of them is *a feeling of anger*, which was not included in the number of occurrences.

The results (see Table 5 and Fig. 1) indicate that the two genres, 'General Fiction' and 'Face-to-face Conversation' correspond most closely in terms of the use of the four forms.[29] All four forms are most frequently used in LOBJ. Form (b) has the lowest frequency among the four forms in all four genres.[30] In terms of grammatical functions, form (c) is rarely used as subject or complement.

Table 5 Occurrences of Forms according to Genres

| | Total words | (a 1)<br>*-tion* + *of* + O | (a 2)<br>Vroot + *of* + O | (b)<br>*-ing* + *of* + O | (c)<br>*-ing* + O | Total |
|---|---|---|---|---|---|---|
| (J) Academic prose | 160,000 | 584<br>(3.65) | 149<br>(0.93) | 110<br>(0.69) | 573<br>(3.58) | 1,416<br>(8.85) |
| (A) Press reportage | 88,000 | 138<br>(1.57) | 41<br>(0.47) | 28<br>(0.32) | 220<br>(2.50) | 427<br>(4.86) |
| (K) General fiction | 58,000 | 25<br>(0.43) | 24<br>(0.41) | 5<br>(0.09) | 123<br>(2.12) | 177<br>(3.05) |
| (S 1,2) Conversation | 140,000 | 63<br>(0.45) | 28<br>(0.20) | 17<br>(0.12) | 313<br>(2.24) | 421<br>(3.01) |
| Total | 446,000 | 810<br>(1.82) | 242<br>(0.54) | 160<br>(0.36) | 1,229<br>(2.76) | 2,441<br>(5.47) |

( ) : Occurrences per 1,000 words

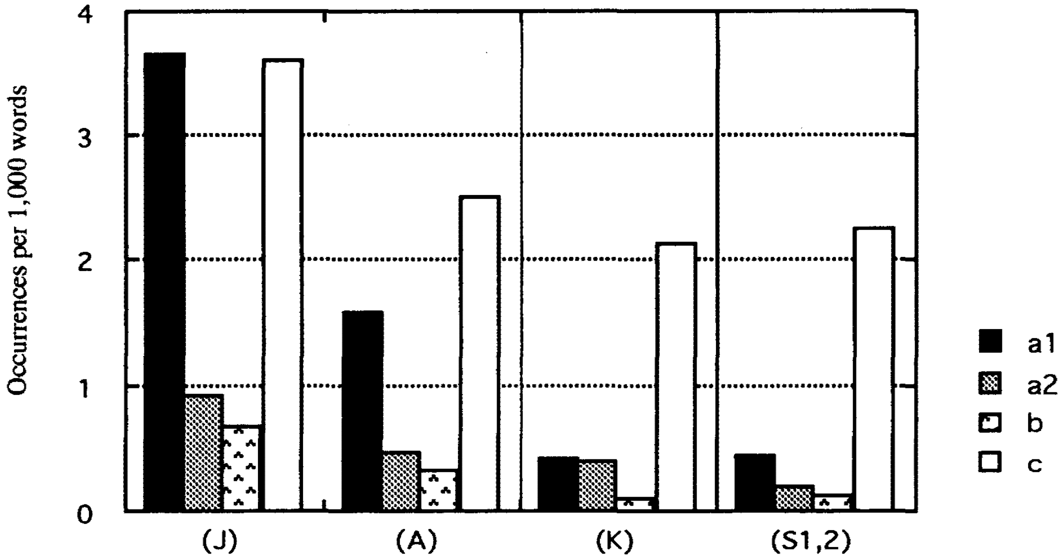


**Fig. 1** Occurences of forms according to genres

(See Fig.2.) The last two results are consonant with trends in early modern English. (Umesaki, 1994 a)

### 4.2 Transitive Verb Nominalisations with *-ing* Suffixes

Quantitative and also qualitative analysis was carried out on the use of form (b) *-ing* + *of* + O. Occurrences of the form counted in all genres of the LOB and London-Lund Corpora show that form (b) is less frequently used in imaginative prose and the spoken genre, and frequently used in informative prose. (See Fig. 3.) Out of 514 instances of form (b), 45 instances occur in which the *-ing* derivative noun has another derivative noun form.

Qualitative analysis indicates that form (b) is used : (1) when the equivalent noun with a *-tion/ment*/etc. suffix (e.g. *destruction*) does not denote 'action' ; (2) to focus on 'action' rather than 'result' (i.e. aspectual difference) ; (3) in accordance with the tendency of *-ing* derivative nouns to take

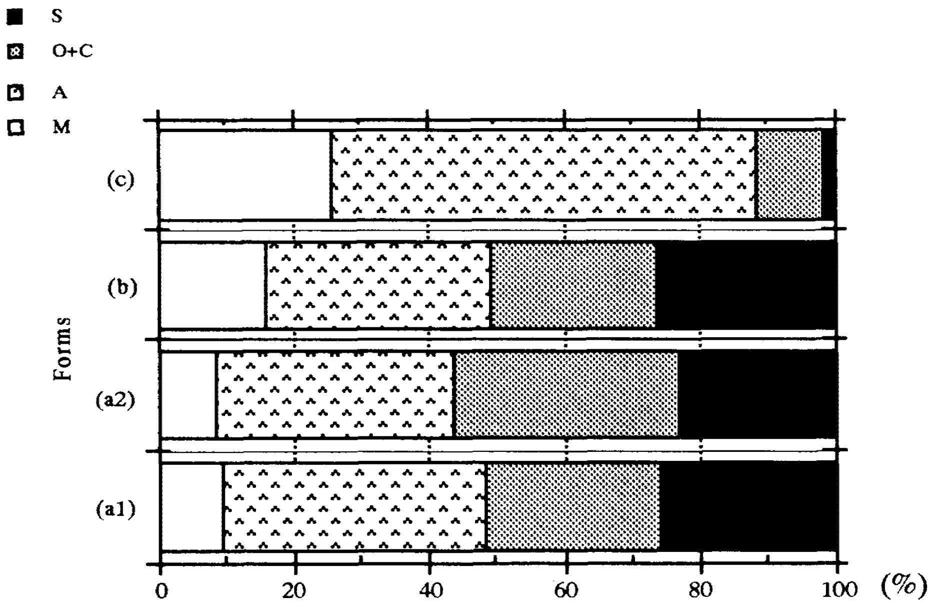


**Fig. 2** Function

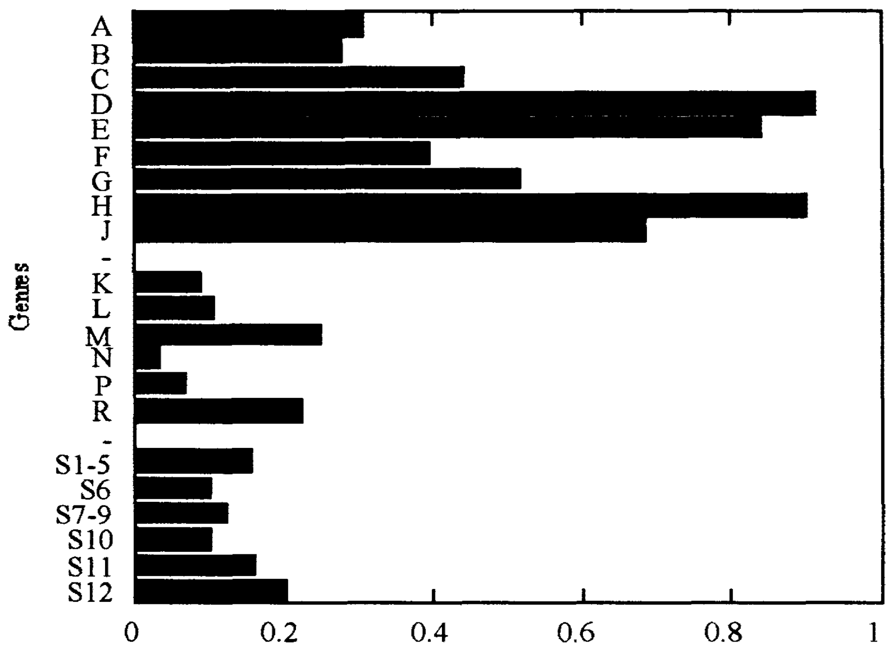


**Fig. 3** Occurrences of *ing* + O according to Genres

*of*-prepositional phrase in verb-object relationship and of nouns with other suffixes to take it in verb-subject relationship ; (4) for discoursal reasons. (Umesaki, 1994 b, 1995 a)

### 4.3 Transitive Verb Nominalisations : Based on Questionnaires

Questionnaires were distributed to British native speakers of English to test hypotheses based on corpus studies. Each question in the questionnaire consists of a sentence containing a target form with contexts extracted from the corpora. The hypotheses tested were : (1) Form (a)-*tion/ment*/etc* + *of* + O tends to be used instead of form (b)-*ing* + *of* + O ; form (b) is chosen when (a) denotes 'result' including a concrete object, when the *of*-prepositional phrase shows a verb-object relationship, when emphasis is placed on an ongoing or incomplete action, and when there are rhetorical or discoursal constraints. (2) Form (a) is preferred to form (c)-*ing*+O in the subject position. (3) Form (c) is preferred to form (b) in the case where form (c) is possible except in the subject and complement position. The hypotheses have been supported by some of the answers given in the questionnaire (excluding 'except in the complement position' in hypothesis (3)). It has been found, however, that informants selected form (a) instead of form (b) where the corpus suggests form (b) is preferred (see Example 1), and that informants chose form (c) instead of (b) in the complement position whereas in the corpus form (b) is selected. This result suggests a tendency towards the avoidance of form (b). This decline of the use of form (b) accords with a historical trend. (Umesaki, 1995 b)

EXAMPLE (1)

It might be said that his lapse in regarding fossils as sports of nature is here offset by his penetration as to their possible use. It would certainly be possible to use a tool of which the true nature was unknown, if, empirically, it had been found to serve a useful purpose. But to credit Lister with the first formulation of the basic principle of stratigraphy, as has been claimed, would be to bestow credit falsely. I think Lister had in mind merely the [**characterizing/characterization**] of different types of rocks by distinctive fossils. Today this would be called recognizing the facies of the rocks and Lister's "ingenious proposal", as it was entitled, to make a map showing the surface distribution of strata was a proposal for a mineral, not a true geological map. (J 02 38)

| corpus : (b) | (b)-*ing* + *of* acc./pref. | (a)-*tion*$^{*}$ + *of* acc./pref. | both acceptable |
|---|---|---|---|
| aged under 50 | 2(1) | 20( 9) | 12 |
| aged over 50 | 5(0) | 12( 3) | 5 |
| Total | 7(1) | 32(12) | 17 |

### 4.4 *To*-Infinitives and Gerunds

In order to find factors which affect the selection of a *to*-infinitive or an *-ing* form as verb complementation of *begin*, *start* and *continue*, the Brown, LOB and London-Lund corpora were used. It has been found that register does not influence the selection of a *to*-infinitive or a gerund as verb complementation (see Table 7). Statistically,[31)] only *begin* in the three verbs showed a significant difference between the Brown and LOB corpora (see Table 6 and Fig. 4). At the morpho-syntactic level, *to*-infinitives are more often used than *-ing* forms after *begin*, whereas both forms are used with relatively similar frequency after *start*. A wider range of verbs in *-ing* can be used after *start* than after *begin*. Semantically, *begin* and *start* followed by *to*-infinitives can be understood as the modality of initiation ; focus is placed not on *begin* and *start* but on the action/event expressed by the non-finite verbs. Where they are followed by *-ing* forms they can be regarded as transitive verbs with a direct object and the focus is placed on the act of initiation. *Continue* is mostly used with *to*-

**Table 6** Verb Complementation Forms of *begin*, *start* and *continue*

| | | Brown | LOB | London-Lund | Total |
|---|---|---|---|---|---|
| *begin* | *to*-infinitive | 258 (45.5%) | 257 (52.6%) | 60 (46.1%) | 575 (48.5%) |
| | gerund | 57 (10.1%) | 24 ( 4.9%) | 6 ( 4.6%) | 87 ( 7.3%) |
| | (vt) noun | 65 (11.4%) | 42 ( 8.6%) | 8 ( 6.2%) | 115 (9.7%) |
| | (vi)–– | 187 (33.0%) | 166 (33.9%) | 56 (43.1%) | 409 (34.5%) |
| | Total | 567 (100%) | 489 (100%) | 130 (100%) | 1186 (100%) |
| *start* | *to*-infinitive | 54 (15.3%) | 36 (10.7%) | 33 ( 9.9%) | 123 (12.0%) |
| | gerund | 57 (16.1%) | 55 (16.4%) | 90 (27.0%) | 202 (19.8%) |
| | (vt) noun | 81 (22.9%) | 87 (26.0%)*1 | 46 (13.8%) | 214 (20.9%) |
| | (vi)–– | 162 (45.7%) | 157 (46.9%) | 164 (49.3%) | 483 (47.3%) |
| | Total | 354 (100%) | 335 (100%) | 333 (100%) | 1022 (100%) |
| continue | *to*-infinitive | 119 (43.4%) | 99 (40.9%) | 17 (41.4%) | 235 (42.4%) |
| | gerund | 5 ( 1.8%) | 11 ( 4.6%) | 2 ( 4.9%) | 18 ( 3.2%) |
| | (vt) noun | 37 (13.5%) | 36 (16.1%) | 7 (17.1%) | 80 (14.5%) |
| | (vi)–– | 113 (41.3%) | 93 (38.4%) | 15 (36.6%) | 221 (39.9%) |
| | Total | 274 (100%) | 239 (100%) | 41 (100%) | 554 (100%) |

*1 includes two *what*-clauses (e.g. : *so he started instead what he called a country club for the rich bwanas. . .* (LOB K 29 : 73)).

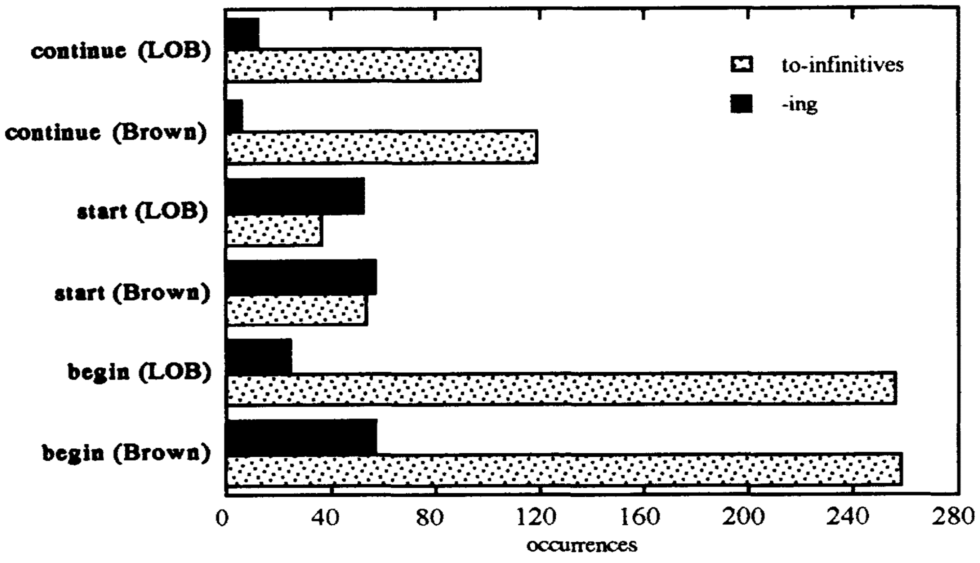


**Fig. 4** Comparison of LOB and Brown

infinitives, entailing the modality of duration. An *-ing* form after *continue* is considered to emphasise the progressive aspect of the action. (Umesaki, 1995 c.)

### 4.5 Nominalisations in Speech and Writing : Analysis Based on a Personal Corpus

In order to investigate the cause of nominalisation in relation to the difference of speech and

**Table 7** Occurrences of *begin, start, continue* according to register

| | | Informative prose<br>(748,000 words) | Imaginative prose<br>(252,000 words) | Speech<br>(500,000 words) |
|---|---|---|---|---|
| *begin* | *to* | 122 (1.63) | 135 (5.36) | 60 (1.20) |
| | *-ing* | 8 (0.11) | 16 (0.63) | 6 (0.12) |
| *start* | *to* | 18 (0.24) | 18 (0.71) | 33 (0.66) |
| | *-ing* | 28 (0.37) | 27 (1.07) | 90 (1.80) |
| *continue* | *to* | 90 (1.20) | 9 (0.36) | 17 (0.34) |
| | *-ing* | 7 (0.09) | 4 (0.16) | 2 (0.04) |

*Values for informative and imaginative prose refer to the LOB Corpus ; speech refers to the London-Lund Corpus.

**Values in round brackets refer to occurrences per 10,000 words.

writing, a corpus was personally compiled. A pair of spoken and written texts consist of an oral presentation at an international conference in the field of natural science (recorded and transcribed by the present writer) and a written paper under the same title by the same scientist as the oral presentation. Three pairs were compiled in a machine-readable form (about 9,000 words of spoken texts and 11,000 words of written texts).

First of all, the oral presentations and written papers were compared by counting the occurrences of 28 linguistic items which Biber (1988) had counted in all the text categories of the LOB and London-Lund Corpora[32]. Cluster analysis was carried out (see Fig. 5) and it was found that the oral presentations and written papers were syntactically distinct. Lexical density is higher in the written texts than in the spoken texts, and it results from the higher frequency of nouns and adjectives modifying nouns in writing.

Three pairs of oral presentations and published papers were compared in terms of how a whole text is structured. It was found that the spoken texts have the summative part (i.e. conclusion) at the end of the speech whereas the written texts have the equivalent part (i.e. an abstract) at the beginning of the paper. This seems to indicate that the most impressive part of the written discourse lies at the beginning whereas that of the spoken discourse lies at the end. The total number of functional components is larger in the written texts than in the spoken texts. With regard to similarities in the oral presentations and papers, it was found that the functional components existing in the development parts of both the texts appear in the same order. It may be suggested that the differences between speech and writing in discourse structure derive from basic differences in the nature of speech and writing, such as those involving reciprocity and temporariness.

Comparison of thematic organisation reveals that in both the spoken and written texts about 70% of the themes appear as subject of a clause. However, the themes of the spoken texts consist far more often of first- and second-person pronouns than those of the written texts. Themes of the writ-

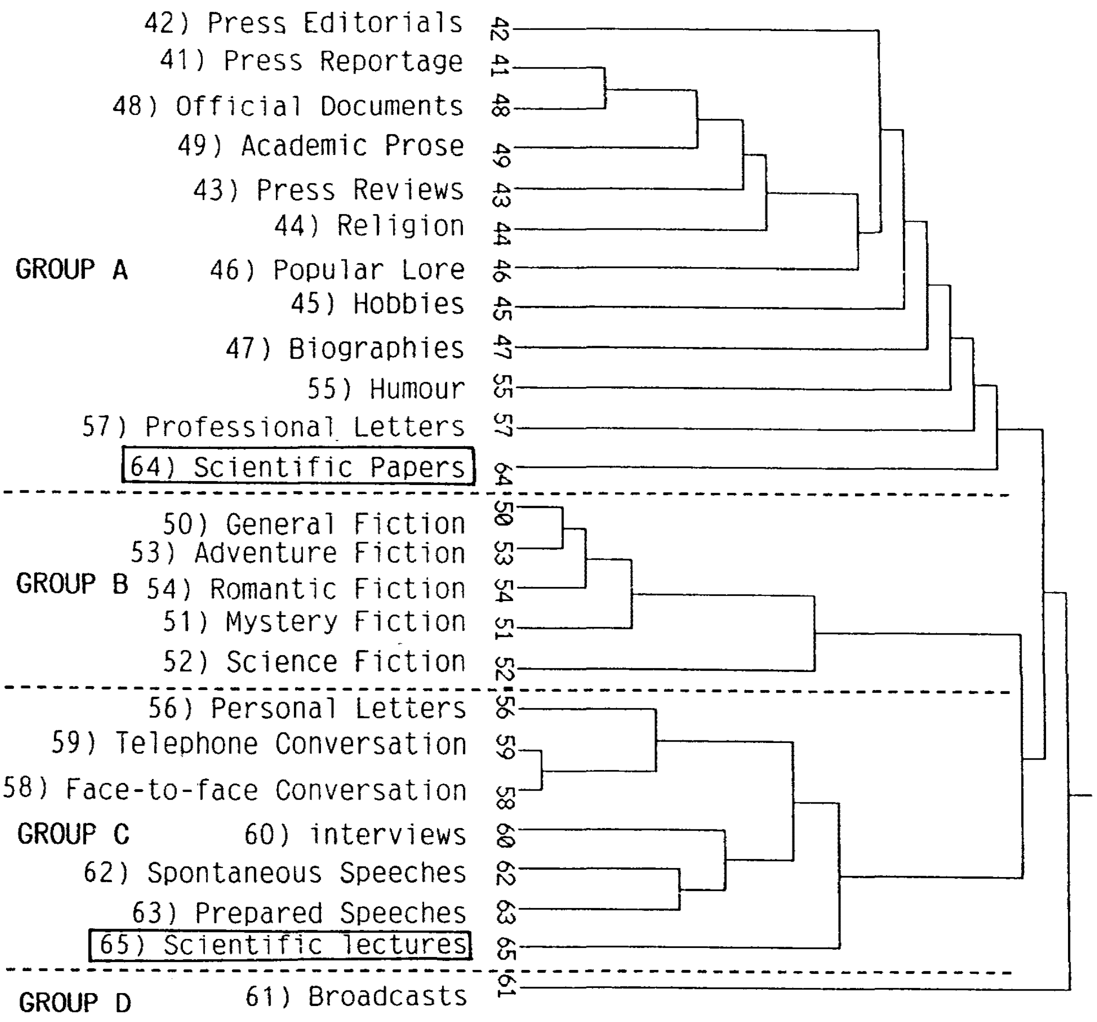


**Fig. 5** Cluster Analysis

ten texts tend to be tied more strongly by lexical cohesion than those of the spoken texts ; and additive conjunctions are much more often used in the spoken discourse than in the written discourse. Noun phrases have played an important role in effecting the lexical cohesion. Initial themes of the functional components in the spoken discourse mostly consisted of *now, so, and*, and/or pseudo-cleft *what*-clauses whereas in the written discourse noun phrases associated with functions of the components are sometimes used in the thematic position.

Direct comparison of speech and writing has shown the important role of noun phrases in the construction of written texts. It can be concluded that in order to organise written texts which greatly depend on noun phrases, the nominalisation of verbs is very important. (Umesaki 1991, 1992 and 1993)

## 5. ADVANTAGES AND LIMITATIONS OF CORPORA

From what has been done so far, some of the possibilities for research provided by corpora and some of the problematic aspects of their use have been identified. The advantages conferred are outlined in what follows, together with suggestions for overcoming difficulties.

(1) A large quantity of examples can be obtained in a shorter time from machine-readable texts than by extracting examples from books manually. It would have been almost impossible to extract the same number of examples as were obtained for 4.1 and 4.4 without corpora and computers. It should be noted, however, that the search for syntactic items (such as nominal *-ing* clauses, *-tion/ment*/etc+ *of*+O), which requires manual procedure, takes a considerable amount of time and energy in comparison with the search for words (such as *begin, start* and *continue*). Improvement of the automatic tagging system, which can, for example, distinguish between gerunds and participles in *-ing* form, might lessen the burden of manual work. Even so, the final judgement in the classification of examples in any analysis will have to be made by researchers.

(2) Multi-genre sample corpora such as the LOB and London-Lund Corpora have made it possible to analyse texts of various categories at the same time and to investigate such linguistic features as register. (See 4.1, 4.2 and 4.4.)

(3) Thanks to the LOB and Brown Corpora, which employ the same organisation of text categories, comparison of British and American English has become possible. It has also been found that statistical analysis functions effectively in discriminating between them. (See 4.4)

(4) The analyses in 4.2 and 4.4 show that linguistic descriptions in previous studies and dictionaries can be checked by examining the corpus qualitatively and quantitatively.

(5) Multi-genre sample corpora can function as a scale for understanding the characteristics of a personally made corpus by the use of Biber's data (1988). (See 4.5.)

(6) It is difficult for a non-native speaker to judge the acceptability of alternatives found in the corpus example. However, questionnaires given to native speakers are effective in establishing acceptability (see 4.3). For the composition of the questionnaire, corpora can provide contextual matter for the forms under consideration.

(7) The Brown, LOB and London-Lund Corpora consist of a fixed number of words per text (2000 words for the first two, and 5000 words for the last) and so it is impossible to study the organisation of a whole text. To compile a corpus consisting of complete texts (as proposed in 4.5) will be a way of overcoming this drawback of the multi-genre sample corpora. It is also expected that this limitation will be overcome by a monitor corpus (see 1.2) in the future.

(8) Even if no example of a form is found in the Brown, LOB and London-Lund Corpora, the form may appear in a larger corpus as was the case with *begin being* (see 4.4). The making of massive-scale corpora may contribute to the solution of the problem of size.

## 6. FUTURE PERSPECTIVES FOR CORPUS-BASED STUDIES

Leech (1992, p.105) says that 'corpus linguistics' refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research. It has a different kind of reference from terms such as 'cognitive linguistics' and 'sociolinguistics' and from 'grammar', 'semantics' and 'pragmatics'. However, determining the characteristics of language according to register will in itself be of significance for the description of the language in use.

The most prevalent way corpora serve this purpose in corpus-based studies is in the investigation of how a form is used. According to Widdowson (1978) 'usage' is the function of a linguistic item as an element in a linguistic system and 'use' is its function as part of a system of communication (Richards et al. 1985). Here, usage and the use of forms are considered as factors in linguistic performance rather than in linguistic competence. This means that not only usage but also the use of a form has to be studied to understand overall performance. Information on how the form is used in different registers or different regional varieties will elucidate the function of the form as part of a system of communication. Corpora can contribute to providing optimal data on usage and also the use of a form.

It must be remembered that corpora cannot record all speech and writing. They deal with representative samples. Because of this, examples picked up from the language which researchers see and hear and store manually in the conventional card form are indispensable to the study of language. Native speakers' intuition is another important data source. It is likely, then, that corpora will be used in combination with such data sources.

The study of linguistic universals and the descriptive approach to language are two pillars on which linguistic research rests. Each is essential, and their proper relations make the secure basis for further accomplishments.

### NOTES

I would like to thank Prof. Edward Costigan for checking the English. All remaining errors are, of course, mine. Special thanks go to Mr. Michinao Matsui for the advice on statistical analysis. I am also very grateful to Prof. Katsumasa Yagi for constant and useful discussion on how to use corpora.

1 ) There are basically two approaches to corpora and their use. One is that researchers obtain evidence to support a contention by extracting examples of a target form from corpora. The other is that they discover rules by examining examples of the target form extracted from corpora. These can be regarded as rational and empirical approaches. Leech points out that corpus linguistics focuses more on the latter approach. This seems to be true, but it may also be the case that the two approaches are often combined : a sample collection is carried out based on the findings of previous studies, and new rules are discovered.

2 ) Translated by the present writer.

3 ) Newspapers such as *The New York Times, The Los Angeles Times, The Washington Post, The Guardian*

and so on are sold in CD-ROM. Complete works of Shakespeare and other writers are also sold in CD-ROM or in diskettes.

4) Leech (1991, p. 11) differentiates an archive from a corpus in terms of 'representativeness'. In this sense, complete works of Shakespeare or full texts of newspapers for a period of time, belong to archives and may not be considered 'corpora'.

5) Sinclair (1991, pp. 23–24) defines a sample corpus as a corpus with a classification into genres (15) of printed text, a large number (500) of fairly short extracts (2,000 words), giving a total of around one million words, and a close to random selection of extracts within genres. He gives the Brown and LOB corpora as examples.

6) The term 'general purpose corpora' is taken from Leech (1991, p. 11). He gives the Brown and the Survey of English Usage Corpus (SEU) as examples. He calls the rest of the synchronic sample corpora 'specialised corpora'. Leitner (1992, pp. 49–50) uses 'a general purpose corpus' and 'special purpose corpora' when considering the structure of the International Corpus of English. In the present study, 'special purpose' applies to the purpose for which corpora are used.

7) Leech (1991, p. 11) uses the term 'a domain-specific corpus' with an example of a corpus representing the language of the oil industry. In the present study not only subject-matter but also other linguistic features are taken into account.

8) The distinction of domain-specific corpora and corpora of a social dialect derives from the difference between registers (diatypic varieties) and dialects (dialectal varieties) which Halliday and Hasan (1985, pp. 41–43) define. They define register as a variety of language according to use, and dialect or dialectal variety as a variety of language according to the user. The terms 'domain' and 'social/regional' are used instead of 'register-specific' and 'dialect-specific' in the present study because spoken/written difference of register and standard regional varieties of dialect have already been categorised as independent features and have to be excluded in this sub-categorisation. (See 2.3 for further definition)

9) Made available through the ICAME (the International Computer Archive of Modern English in the Norwegian Computing Centre for the Humanities : Harald Haarfagresgate 3 N-5007 Bergen Norway ; icame @hd.uib.no.). The CD-ROM can be used in IBM, MAC and UNIX machines.

10) With regard to definition see 2.3.

11) This table was adapted from Akano, I. and K. Fujimoto (1994, p. 25).

12) The Brown Corpus is also grammatically tagged, but the tagged version is not distributed by the ICAME.

13) The meaning of each symbol is listed in the table given in Svartvik et al., eds. (1982, pp. 26–28)

14) Some other corpora available through the ICAME are : (1) the Polytechnic of Wales Corpus (Ca. 61,000 words of children's speech excluding those with strong second language influence, Welsh or other) ; (2) the Melbourne-Surrey Corpus (100,000 words taken from the newspaper *The Age* published in Melbourne from Sept.1, 1980 to Jan. 30, 1981) ; (3) the Lancaster Parsed Corpus (133,000 words of syntactically analysed sentences taken from each text category of the LOB corpus) ; (4) the Lancaster/IBM Spoken English Corpus. All are in diskettes of MS-DOS : 720 KB, 1.2 MB, 1.4 MB, MAC : 800 KB, 1.4 MB.

Those published by Oxford University Press are : (1) MicroConcord Corpus Collection A, (One million words taken from the British newspapers *The Independent* & *The Independent on Sunday*, October/November 1989) ; (2) MicroConcord Corpus Collection B, (One million words taken from academic titles published by Oxford University Press) ; both in five 1.4 MB diskettes.

15) Accessible categories are American books (2.5 M), Ephemera (1.8 M), Mags (2.6 M), American radio (2.7 M), British books (2.5 M), Spoken (3.2 M), *Times* newspaper (2.7 M), *Today* newspaper (2.5 M).

e-mail : direct@cobuild.collins.co.uk

for trial : telnet 193.112.240.130

(type 'codemo' after 'login', type 'cobdemo' after 'cobdemo's Password')

mailing address : COBUILDDirect, Institute of Research and Development, Birmingham
Research Park, Vincent Drive, Birmingham B 15 2 SQ, England

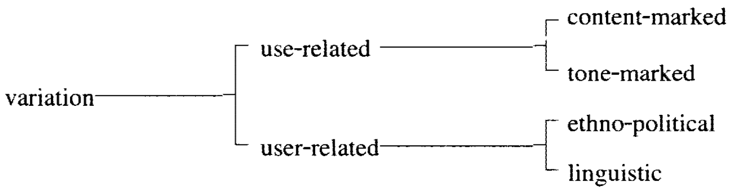
fax : 44–121–414–6203 (COBUILDDirect)

16) Johannson (1991, p. 312) gives the number of publications based on or related to the English text corpora distributed through ICAME :

| | Year | Number of publications |
|---|---|---|
| | -1965 | 10 |
| 1966 | -1970 | 20 |
| 1971 | -1975 | 30 |
| 1976 | -1980 | 80 |
| 1981 | -1985 | 160 |
| 1986 | - | 320 |

17) For example, see Marutani et al. (1994).

18) Quirk et al. (1990, pp. 48–53) recognise two factors in language varieties : use-related and user-related. These two factors are considered to be equivalent to varieties according to use and those according to user in Leech et al. (1982, pp. 6–10), Halliday (1978, p. 35) and Halliday et al. (1985, pp. 39–43). However, the following sub-categorisation given by Quirk et al. differs from Halliday's sub-categorisations of register and dialect. (See NOTE 7).



19) Other definitions are given by Leckie-Tarry (1995, pp. 7–9) from the viewpoint of systemic functional grammar, and Coulthard (1985, pp.40–42) in relation to stylistic structures.

20) The following package programs can output the KWIC concordance together with other data.

*MicroOCP* : available through Okita Denshi for PC 9801, through Oxford University Press for IBM PC. KWIC concordances, word lists, frequency lists and word index are obtained. (Reference : Nagase, M. and H. Nishimura, 1986 ; Akano, I., 1990)

*WordCruncher* : available through ICAME or Johnson & Company (778 South 400 East, OREM, Utah, 84058, U. S. A.) for IBM PC. This package program can index DOS text files, and by the use of the indexed text files, KWIC concordances and word indexes. Statistical data such as z-scores and frequency distributions are also obtainable.

*TACT* (Text Analysis Computing Tools) : available through ICAME for IBM PC or via Internet (FTP> epas. utoronto. ca.). This program can help with encoding texts, make a KWIC concordance and calculate word and collocated word frequencies and make statistical analysis.

*MicroConcord* : available through Oxford University Press for IBM PC. This package program can produce a KWIC concordance, which is further used to find frequency of collocated words and so on. (Reference : Inoue, N., 1995)

*Free Text Browser* : available through ICAME for Macintosh personal computers. This program enables us to index text files and make a concordance, which is further processed to find frequency of collocated words and so on. To use this program, *HyperCard* is necessary. (Reference : Yoshimura, Y. 1994)

*Concorder :* available through University of Montreal (C. P. 6128-A, Montreal, Quebec H 3 C 3 J 7) for Macintosh personal computers. Texts not only in English but also in other languages can be processed.

*egrep* : available via e-mail (NIFTY serve). A package program with other functions than KWIC concordance such as lemmatisation has become available.

*Lexa* : available through ICAME for IBM PC or via Internet (ftp>nora.hd.uib.no (129.177.24.42) for further detail, see *ICAME Journal* 19, pp. 164–165). This program can lemmatise texts, create frequency lists of the types and tokens, make lexical density tables and so on.

21) There are package programs with which each example is stored in a card form, for example *The Card* for PC 9801, *HyperCard* for Macintosh personal computer.

22) *VZ Editor* (version 1.0, copyright : Village Center) for PC 9801.

23) JGAWK (version 1, distributed by Free Software Foundation) for PC 9801.

24) For reference, Fasold, R. (1984) pp. 85–146 ; Woods, A, P. Fletcher and A. Hughes (1986) ; Butler, C. (1985).

25) *SRI Tokei Package* (version 2.5, copyright : Social Research Information Service Inc., Tokyo) for PC 9801.

*Statistica*(copyright : Threes Company Inc.) for Macintosh.

*SPSS* (available through SPSS Japan Inc.) for IBM, PC 9801, Macintosh and UNIX.

*CaleidaGraph* (version 2.1.3, copyright Abelbeck Software) for Macintosh.

26) The machines used for the automatic input of scripts for the present analysis is a Kurzweil K 5100 optical character reader and a Sanyo AY 386 personal computer.

27) Lees (1960) uses the term 'nominals' instead of 'nominalisations'.

28) Quirk et al. (1985, p. 1063 and pp. 1290–1292), who reject the use of the term gerund, include appositive (*His current research, investigating attitudes to racial stereotypes, takes up most of his time*) and adjectival complementation (*They are busy preparing a barbecue*) in their listing of functions of nominal *-ing* clauses.

29) The F-test carried out by the use of the values obtained from the Analysis of Variance (ANOVA) supports this result. In the following table, values larger than 1.4126 indicate that the two genres compared are significantly different in respect of the use of the forms ; those smaller than 1.4126 that there is no significant difference. The smallest value indicates closest similarity.

l ij=1.4126

| | J | A | K | S |
|---|---|---|---|---|
| J | | 1.0000 | 1.4525 | 1.4625 |
| A | | | 0.4525 | 0.4625 |
| K | | | | 0.1000 |
| S | | | | |

30) The Chi-square test supports this result. With form (b) as model, the values of (a 1) (4.402) and (a 2) (6.550) are not significant at the level of 0.05, whereas the value of (c) (43.953) is significant at the level of 0.005.

31) The Chi-square test indicates a significant difference between the Brown and LOB Corpora in the case

of *begin* ($T=11.54>x^2(1, 0.025)=5.024$).

32) The 28 linguistic items are grouped into 13 categories and the following 11 are used for cluster analysis : conjuncts, passives, nominalisations, participles (post-modification), gerunds, participles (adverbial), infinitives, relatives, adverbial subordinators, *that*-complementations, coordinators.

## BIBLIOGRAPHY

Aijmer, K. and B. Altenberg, eds. (1991) *English Corpus Linguistics*, Longman, London.

Akano, I. (1989) "Corpus Linguistics eno shotai (1) — Brown Corpus —," *SELL* 6, 142–148.

Akano, I. (1990) "Corpus Linguistics eno shotai (2) — OCP toha nanika —," *Journal of Kyoto University of Foreign Studies* 35, 1–15.

Akano, I. (1995) "Kopasu Gengogaku no Saikin no Dokou," handouts presented at the July meeting of Rokko Eigogaku Kenkyukai, unpublished.

Akano, I. and K. Fujimoto (1994) "Participial Constructions in Corpora," *English Corpus Studies* 1, 19–34.

Akano, I., Y. Yoshimura and K. Fujimoto (1991) "Corpus Linguistics no genzai no doukou to mondaiten (1) kopasu to sono kochiku," *SELL* 7, 1–45.

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge University Press, Cambridge.

Butler, C. (1985) *Statistics in Linguistics*, Basil Blackwell, Oxford.

Butler, C., ed. (1992) *Computers and Written Texts*, Basil Blackwell, Oxford.

Clear, J. (1992) "Corpus Sampling," in Leitner, G., ed. (1992) 21–31.

Coulthard, M. (1977/1985) *An Introduction to Discourse Analysis*, Longman, London.

Crystal, D. (1987) *The Cambridge Encyclopedia of Language*, Cambridge University Press, Cambridge.

Fasold, R. (1984) *The Sociolinguistics of Society*, Basil Blackwell, Oxford.

Garside, R., G. Leech, and G. Sampson, (1987) *Computational Analysis of English*, Longman, London.

Greenbaum, S. (1992) "A New Corpus of English : ICE," in Svartvik, J., ed. (1992) 171–179.

Halliday, M. A. K. (1978) *Language as Social Semiotic*, Edward Arnold, London.

Halliday, M. A. K. and R. Hasan (1985) *Language, Context, and Text : Aspects of Language in a Social-Semiotic Perspective*, Deakin University, Victoria.

Inoue, K., H. Yamada, T. Kono and H. Narita (1985) *Gendai no Eibunpou 9, Meishi*, Kenkyusha, Tokyo.

Inoue, N. (1994) "Cobuild Corpus : The Bank of English to wa Nanika," *Eigo Kyoiku to Eigo Kenkyu* 11, *Shimane University.*

Inoue, N. (1995) "MicroConcord — Concordance Program," (in Japanese) *English Corpus Studies* 2, 141–148.

Johansson, S. (1991) "Times change, and so do corpora," in Aijmer, K. and B. Altenberg (1991) 305–314.

Johansson, S., E. Arwell, R. Garside and G. Leech (1986) *The Tagged LOB Corpus, Users' Manual*, Norwegian Computing Centre for the Humanities, Bergen.

Johansson, S. and Stenström (1991) *English Computer Corpora, Selected Papers and Research Guide,* Mouton de Gruyter, Berlin.

Kytö, M., M. Rissanen and S. Wright, eds. (1994) *Corpora across the Centuries*, Rodopi, Amsterdam.

Leckie-Tarry, H. (1995) *Language & Context*, Pinter Publishers, London.

Leech, G. (1991) "The State of the Art in Corpus Linguistics," in Aijmer, K. and B. Altenberg, eds. (1991) 8–29.

Leech , G. (1992) "Corpora and Theories of Linguistic Performance," in Svartvik, J., ed. (1992) 105–122.

Leech, G., M. Deuchar and R. Hoogenraad (1982) *English Grammar for Today*, Macmillan, London.

Leech, G., and S. Fligelstone (1992) "Computers and Corpus Analysis," in Butler, C., ed. (1992) 115–140.

Leech, G., R. Garside and M. Bryant (1994) "The Large-Scale Grammatical Tagging of Text : Experience with the British National Corpus," in Oosdijk, N. and P. de Haan, eds. (1994) 47–63.

Lees, R. B. (1963) *The Grammar of English Nominalizations*, Mouton, The Hague.

Leitner, G. ed. (1992) *New Directions in English Language Corpora*, Mouton de Gruyter, Berlin.

Marutani, M., Y. Higuchi, A. Inagi, N. Takao, S. Ando, M. Fujimoto (1994) "Corpus Processing Instruction : The Case at Otemon Gakuin University," (in Japanese), *English Corpus Studies* 1, 63–84.

Nagase, M. and H. Nishimura (1986) *Konpyuta ni yoru Bunshokaiseki Nyumon : OCP eno Shotai*, Ohmusha, Tokyo.

Nakamura, J. (1993) "Quantitative Comparison of Modals in the Brown and the LOB Corpora," *ICAME Journal* 17, 29–48.

Nakamura, J. (1994) "Recent Trends in Corpus-Based English Studies in Europe," (in Japanese), *English Corpus Studies* 1, 49–61.

Oostdijk, N. (1991) *Corpus Linguistics and the Automatic Analysis of English*, Rodopi, Amsterdam.

Poutsma, H. (1926) *A Grammar of Late Modern English, Part II, the Parts of Speech II*, P. Noordhoff, Groningen.

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

Quirk, R. and G. Stein (1990) *English in Use*, Longman, London.

Richards, J., J. Platt and H. Weber. (1985) *Longman Dictionary of Applied Linguistics*, Longman, London.

Saito, T., ed. (1992) *Eigo Eibungaku Kenkyu to Konpyuta*, Eichosha, Tokyo.

Saito, T. (1994) "Eigo-kopasu Kenkyu no Saikin no Doukou," *The Rising Generation*, 14–16.

Sinclair, J. (1991) *Corpus Concordance Collocation*, Oxford University Press, Oxford.

Svartvik, J., ed. (1992) *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*, Mouton de Gruyter, Berlin.

Svartvik, J., M. Eeg-Olofsson, O. Forsheden, B. Orestroem, and C. Thavenius (1982) *Survey of Spoken English*, Wallin & Dalholm Boktr AB, Lund.

Swales, J. (1981 a) *Aspects of Article Introductions*, The University of Aston, Birmingham.

Umesaki, A. F. (1991) "A Comparison of Scientific Lectures and Papers with regard to Grammar and Lexis : Differences between Speech and Writing," *Studies in Modern English* 9, 37–53.

Umesaki, A. F. (1992) "An Analysis of Spoken and Written English based on Halliday's Theory of Lexical Density," (in Japanese) *Osaka University Journal of Language and Culture* 1, 113–123.

Umesaki, A. F. (1993) "Thematic Organisation in Academic Speech and Writing," *Aspects of Modern English*, 470–487, Eichosha, Tokyo.

Umesaki, A. F. (1994 a) "A Corpus-Based Study of Transitive Verb Nominalisations : Nominal *-ing* Clauses and Derivative Noun Phrases," *English Corpus Studies* 1, 85–98.

Umesaki, A. F.(1994 b) "The Use of Transitive Verb Nominalisations with *-ing* Suffixes in Speech and Writing : A Corpus-Based Study", *Helicon* 19, 35–48.

Umesaki, A. F.(1995 a) "A Corpus-Based Study of the Selection of Nominalised Forms," *Journal of Tezukayama College* 32, 95–108.

Umesaki, A. F. (1995 b) "A Study on the Use of Nominalisations in English : Based on Questionnaires," *Helicon* 20, 95–120.

Umesaki, A. F. (1995 c) "A Corpus-Based Study of *To*-Infinitives and Gerunds as Verb Complementation : *Begin, Start* and *Continue*," *The JASEC Bulletin* 4, 57–66.

Widdowson, H. (1978) *Teaching Language as Communication*, Oxford, Oxford University Press.

Widdowson, H. (1983) *Learning Purpose and Language Use*, Oxford University Press, Oxford.

Woods, A, P. Fletcher and A. Hughes (1986) *Statistics in Language Studies*, Cambridge University Press, Cambridge.

Yagi, K. (1987) *Atarashii Goho Kehkyu*, Yamaguchi Shoten, Kyoto.

Yoshimura, Y. (1995) "Free Text – Macintosh yo Kensaku Puroguramu –," *English Corpus Studies* 2, 149–154.