Using Corpora for Studying English Grammar

Heok-Seung Kwon
(Seoul National University)

Kwon, Heok-Seung. 2004. **Using Corpora for Studying English Grammar**. *Korean Journal of English Language and Linguistics* 4-1, 61-81. This paper will look at some grammatical phenomena which will illustrate some of the questions that can be addressed with a corpus-based approach. We will use this approach to investigate the following subjects in English grammar: number ambiguity, subject-verb concord, concord with measure expressions, and (reflexive) pronoun choice in coordinated noun phrases. We will emphasize the distinctive features of the corpus-based approach, particularly its strengths in investigating language use, as opposed to traditional descriptions or prescriptions of structure in English grammar.

This paper will show that a corpus-based approach has made it possible to conduct new kinds of investigations into grammar in use and to expand the scope of earlier investigations. Native speakers rarely have accurate information about frequency of use. A large representative corpus (i.e., The British National Corpus) is one of the most reliable sources of frequency information. It is important to base an analysis of language on real data rather than intuition. Any description of grammar is more complete and accurate if it is based on a body of real data.

Key Words: Corpora, English grammar, number, concord, reflexive pronoun

1. Introduction

Computers have made it possible to identify and analyze complex patterns of language structure and use, allowing the analysis of a much larger database of natural language than could be dealt with manually. According to Sinclair (1991, p. 36),

the discrepancies between the way a computer can identify elements in a text and the expectations of a linguist who knows the language of the text is well worth investigation. In this paper I will emphasize the distinctive features of a corpus-based approach, particularly its strengths in investigating language use, as opposed to traditional descriptions or prescriptions of structure in English grammar.

Corpus data enables us to see a variety of linguistic aspects. This paper will explore the role of corpus analysis in describing some features of English grammar by examining the frequency of a word or a structure in real language. The British National Corpus (henceforth, BNC),[1] *The Times* (1995) and *Time* (1989-1994) will be used in the analysis. The BNC aims to represent contemporary British English, and the other two collections are used to check whether the results of analysis of different sources of text correspond closely.

This paper will look at some grammatical phenomena which will illustrate some of the range of questions within traditional English grammar that can be addressed with a corpus-based approach. There are many advantages of a corpus-based approach to language analysis. We will use this approach to investigate the following areas in English grammar: number ambiguity, subject-verb concord, and (reflexive) pronoun choice in coordinated noun phrases.

In order to compare language use patterns for the word or structure in question, it is essential to know how many times each word or structure occurs in real language. We will first carry out a quantitative analysis to see how often a certain pattern occurs relative to other patterns in a representative corpus.

---

[1]The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.

2. Lexis: Number Ambiguity

Should *data* be used as a singular or as a plural noun? The historical answer is clear: the Latin *datum* is singular and *data* is its plural form. Some foreign plurals, however, are variously treated as singular or plural. The Latin plurals *data* and *media* have become dissociated from the original singular forms *datum* and *medium*, respectively, and are variably treated as both plural and singular nouns. With *data*, however, both singular and plural are possible.

In this paper, attention is focused mainly on words borrowed from Latin, especially Latin nouns ending in *-um* or *-on*. Some Latin nouns ending in *-um* keep their foreign plurals (e.g., *acquaria, curricula, media, memoranda, millenia, phenomena, spectra, strata,* etc.),[2] or there may be alternation with the regular plural form (e.g., *forums, stadiums*). In the following sections, some of the more frequently occurring words such as *data* and *criteria* will be examined.

2.1. Dictionary Information

First, consider the word *data*. One method of trying to determine whether *data* is singular or plural is to ask a native speaker. In case native speakers are not available, the second traditional method is to consult a dictionary.

Let us now consider how recent learner's dictionaries deal with the word *data* in their entries:

Longman Advanced American Dictionary (LAAD) (2000)
data n. [plural] information of facts that have been gathered in order to be studied: *All the data shows that these animals are more adaptable than we thought.*

---

[2]According to Biber D. *et al.* (1999), *forums* and *stadiums* are preferred over the irregular forms *fora* and *stadia*. S-plurals are also attested for: *aquariums, curriculums, maximums, memorandums, milleniums, spectrums.*

Cambridge Dictionary of American English (CDAE) (2000)

data *n* [U] information collected for use · *They had data on health, education, and economic development.*

· USAGE: Although originally a plural (the rarely used singular is **datum**) and used with a plural verb, **data** is now often used as an uncount noun with a singular verb.

Cobuild English Dictionary for Advanced Learners (COBUILD) (2001)

data datum is sometimes used as the singular form of **data**. You can refer to information as **data**, especially when it is in the form of facts of statistics that you analyse. In American English, **data** is usually a plural noun. In technical or formal British English, **data** is sometimes a plural noun, but at other times, it is an uncount noun.

...the latest year for which data is available... To cope with these data, hospitals bought large mainframe computers.

Macmillan English Dictionary (MED) (2002)

data noun [u] ★★★

1. facts or information used for making calculations or decisions: can be followed by a plural verb in scientific English, in which case the singular is **datum**: *The analysis was based on data collected in the field.*

2. information in a form that a computer can use: *The new format carries 30 times more data than a CD-ROM.*

LAAD (2000) contains information about the grammar of words. In the entry for the headword *data,* the dictionary indicates that the word is always plural. Despite this the grammatical information given in the brackets does not correspond with the example. The verb *shows* in the example *"All the data shows..."* is clearly in conflict with the grammatical information. The example doesn't help to show that the word *data* is plural. Examples should be chosen to show the ways in which a word or phrase is typically used. However, we cannot exclude the possibility that the example has been deliberately selected to show that the word *data* is used as both plural and singular.

CDAE (2000) indicates that the word *data* is uncountable. Additionally, the dictionary provides a separate note on the

usage of the word. The usage note states that the word takes both singular verb forms and plural verb forms, which is not explained enough for the reader to clearly distinguish between singular and plural.

COBUILD (2001) supplies very specific information about the usage of the word *data*: 'In American English, *data* is usually a plural noun. In technical or formal British English, *data* is sometimes a plural noun, but at other times, it is an uncount noun'. It thus shows that the word *data* is as both singular and plural in different regions and for different purposes. In addition, the two examples are chosen to show the ways in which *data* is used as both singular and plural.

MED (2002) includes specific information about the grammar of the word *data* in the entry: the grammar code [u] means that *data* is an uncountable noun that cannot be used with *a* or *an*. In sense 1, however, the dictionary mentions the use of *data* as plural in scientific English. The dictionary distinguishes between one use and the other, but the two examples provided in the entry for the headword *data* do not clearly show the ways in which *data* is used as either singular or plural.

To summarize, it is shown that while the use of *data* as both singular and plural is explicitly stated in CDAE (2000), COBUILD (2001) and MED (2002), the information is implicitly provided in LAAD (2000). It seems that because of lack of space in a printed dictionary or because of the careless ways in which examples are selected, dictionaries can only give a brief account of the usage of the word in question. Once we start looking at corpus evidence, however, we will be able to see how much more information we can obtain.

2.2. Corpus Information

Now let us look at example sentences that show the ways in which the word *data* is used as both singular and plural. The

following examples are taken from the BNC:

1. It's simple! **These** data must be available.
2. **These** data provide a plausible explanation for the French paradox.
3. Taken together, **these** data indicate that a DNA-dependent kinase extracts...
4. Clearly, **these** data lend strong support to the reaction mechanism outlined in Fig. 1.
5. To our knowledge, **these** data are the first evidence for a specific interaction...
6. Because eurobonds are bearer instruments, data **are** not available on the investor base.
7. Instead, data **are** provided directly and more timely to obviate this need.
8. On the next pulse, data **is** transferred from D1 (1) to Q1 so Q1 = 1...
9. Data **is** aggregated into pairs of adjacent risk ratings...
10. In this chapter, data **is** presented from four of the questions,
11. Data **is** transferred at 33Mb a second between the 32-bit EISA bus...
12. Data **is** written to a floppy disk by a double set of heads that...
13. **This** data will be analysed for any patterns or trends of activity.
14. **This** data can be subjected to the student's t-test for statistical significance.
15. **This** data is plotted on the graph and a line is drawn to link the points.
16. Viewed in isolation, **this** data appears meaningless (see Screen 2)...
17. The data **is** explicitly destined for public use.
18. The data **are** postcoded and may have some occupational details.
19. However, when the data **are** studied by a group of people who...
20. The point to emphasise is that the data **is** only of help to those who know the patient.

Given the large number of cases in which *data* is used as either plural or singular, it is possible to think that there is a distinction between one use and the other, as is stated in some dictionaries. On closer inspection, however, these examples show that in many cases no clear distinction is being made between

plural use and singular use. The word *data* is very frequently used, but people seem to be unsure whether it is a singular or a plural word.

Note incidentally that the word *data* seems to be more frequently used as a plural word, and this is reflected in the evidence from *Nature, New Scientist,* and *The Guardian.* Jones (1997, Kibbitzer 7) provides statistical information about frequencies of the word *data* used as either a singular word or a plural word.[3] Table 1 shows the frequency of *data* in each publication: the proportion of citations in which *data* is unambiguously marked as plural or singular: and, of those, the proportion in which it is marked as plural and the proportion in which it is marked as singular.

Table 1

Frequencies of *data* in *Nature, New Scientist, The Guardian*

(Source: Jones, Kibbitzer 6)

publication	total number of occurrences	plural	singular
Nature	187	175 (93.6%)	12 (6.4%)
New Scientist	123	84 (69.9%)	39 (30.1%)
The Gurdian	81	32 (39.5%)	49 (60.5%)

The three publications show a marked difference between the use of *data* as a plural or singular. In the high-level science journal *Nature,* the use of *data* as a plural is dominant (93.6%). In the more popular science journal *New Scientist,* the use of *data* as a singular form is more frequent, though it is still outnumbered by its use as a plural (69.9%). In educated English usage as represented by the *Guardian* newspaper, *data* is less frequently used as a plural (39.5%). The data in Table 1

[3]For details about frequencies of the word *data,* refer to the website (http://web.bham.ac.uk/johnstf).

demonstrates that in total, the word *data* is treated as a plural word (291 instances) three times as often as a singular word (100 instances).

We can conclude from this statistical information that while in a technical or scientific context *data* is more frequently treated as a plural word, in a non-scientific context it is more often treated as a singular word. It is evident from corpus analysis that using a large corpus of natural language provides better information about the usage of the word in question than any other source.

2.3. Is *criteria* on the Way Towards Singularity?

It is worth noting here that the word *criteria* is regularly used with plural concord, but with the odd exception. This word seems to behave in a way like the word *data* considered in the previous section. This section looks more closely at the singular use of the word *criteria.*

Singular use of *criteria* occurs in the BNC. Of the 3,935 occurrences of *criteria,* there are 23 examples of singular uses such as *one criteria, the first criteria is, criteria has become,* and *I have one other criteria.* Let us now look at more examples taken from the BNC and *The Times* (1995):

<23 occurrences in the British National Corpus>

1. ...York on the basis of **one** criteria, even if that was **a** valid criteria in the terms...
2. The criteria which we'll come on to debate there is **one** criteria in there...
3. There was only **one** criteria for entry into the scheme...
4. That's **one** criteria. And then it's got to be regional, national or international...
5. There has only ever been **one** criteria for choosing a Wedding Present single.
6. ...in small quantities and will be used more now the design criteria **has** been agreed.
7. ...because criteria **has** become one of the central issues of the current educational debate.

8. ...as you're moving south in the new situation **a** new criteria is production.
9. The criteria **is** I think, will somebody give us the funding to do this?
10. ...that's no criteria **is it**? No but why, how parents can let Yeah.
11. Well you see this is where basically the criteria **is** what is the most...
12. The criteria **is** that the songs have to have character.
13. ...they got two it depends on what their criteria **is** if it's...
14. I mean it's 56767 numbers, that's all the criteria **is**.
15. The most important criteria **is** for patients and clients to be able to...
16. What criteria **is** the court to apply?
17. If the same criteria **is** used however, as is used to...
18. Obviously the first criteria **is** to make a profit...
19. Broadly the criteria **is**: has there been an enhancement of the future benefits...
20. The final criteria **is** applied to the valuation of the business...
21. ...open shows such as the Whitechapel whose selection criteria **is** open to artists...
22. The first criteria **is**, what? A clear and precise remit is obtained and documented .
23. Erm I have **one** other criteria which I would suggest you'd need to take into account...

<5 occurrences in *The Times* (1995)>

1. Detailed criteria **has** been set out as to what is meant by competence.
2. What other criteria **is** there besides ability?
3. Our main criteria **is** to create products based on or inspired by Trust property.
4. Most relied on parents' geographical proximity to the school as at least **one** criteria...
5. When my husband and I took out our mortgage 31 years ago, the criteria **was** 2.5 times his salary and that was it.

Of 3,935 occurrences of *criteria* in the BNC, 23 (0.6%) are marked as being singular.[4] Of 804 occurrences of *criteria* in *The Times*

[4]There are *one other criteria* (1), *one criteria* (5), *criteria has* (2), and *criteria is* (15). The number in the parenthesis refers to the frequency of

(1995), 5 (0.6%) are marked as being singular. The number of the singular use of *criteria* accounts for only 0.6% of the total occurrences of the word in the BNC and *The Times*, respectively, but it seems that they are a good, if not immediate, indication of the conceptual change taking place in the mental lexicon of English speakers. It may be that the word *criteria* (plural of *criterion*) has started along the path taken by *agenda* and on which *data* is already well under way. It will thus be tempting over the next decades to observe whether the word *criteria* continues on the route towards singularity.[5)]

3. Subject-Verb Concord

The subject and the verb phrase in the sentence agree in number and person. With the exception of the verb *be*, this subject-verb concord is limited to the present tense. The basic grammatical rule is that the *s*-form of lexical verbs is used with a third person singular subject in the present tense indicative. In practice, however, concord patterns are not always straightforward. Sometimes the number of the subject may be in doubt, either because it is not clearly marked or because the number of the subject noun phrase is variable. For example, there are complications associated with the form of the subject, the meaning of the subject, and the distance between the head of the subject noun phrase and the verb phrase.

3.1. Concord in the Pattern of [one of the few who...]

The regular pattern of grammatical concord may be disturbed by the principle of proximity, i.e. the tendency for the verb to

---

the structure in the BNC.

[5]It is also interesting to note in this connection that the word *phenomena* is not much different from the word *criteria*. Of the 1,361 occurrences of *phenomena* in the BNC, there are 6 examples of the singular use, which accounts for 0.4% of the total occurrences of *phenomena*.

agree with a noun that is closer to the verb (typically in a postmodifier) but which is not the head of the subject noun phrase. In one particular pattern there is a deviation from grammatical concord which seems to work against the principle of proximity—in relative clauses with an antecedent noun phrase [one of + a plural noun phrase].

In this section, we will look particularly at the pattern of [one of the few who + verb] and [one of the few + noun + who + verb]. Here is an example sentence that worries both teachers and learners of English:

> Jackson is one of the few leaders who **has/have** tough assignments.

Here the problem is whether the verb *have* should be the singular form *has* to agree with the preceding singular subject "Jackson", whether it should be singular to agree with the numeral "one", or whether it should be plural to agree with the preceding plural subject "leaders".

Now let us see which is the preferred type of verb concord in this structure. Biber *et al.* (1999, p. 190) gives a few illustrative examples which clearly raise the issue:

> I realize I am [one of the very few Americans] who **knows** the sound of rocks cutting through flesh and striking one. [FICT]
> Mr Devaty is [one of the few dissidents] who **do** not come from a Prague-based intellectual background. [NEWS]

Swan (1995, p. 529) also addresses the issue with the following example:

> She's one of the few women who **has/have** climbed Everest.

Here the verb *have* is plural, because its subject *who* has a plural

reference *the few women*. However, the verb in the relative clause can also take the singular form *has*, because its subject *who* of the relative clause can have a singular antecedent *one*. In addition, the sentence is also saying that *She has climbed Everest*, and in an informal style many people would therefore say *She's one of the few women who has climbed Everest*.

Although this is not strictly correct (the verb in the relative clause should agree with the subject of the relative clause, not with the subject of the main clause), structures of this kind are very common in English, as can be seen in the following examples:

> One of the things that really **make/makes** me angry is people who don't answer letters.
> Alice was one of the students that **were/was** late for the lecture.

It is not clear whether the verbs (*make/makes, were/was*) should take the singular or plural form in these sentences. The best thing to do is to study real examples in a corpus, which can teach us both about the frequency of the structure in question and about the differing functions of particular variants.

Now let us look at more examples that show both singular and plural number concord. The examples to follow are taken from the BNC, *The Times* (1995), and *Time* (1989-1994). Instances where the verb form does not show number concord (e.g., *Fuchs was one of the few public figures who supported Beeren publicly...*) are omitted.

1. Aspinall is one of the few who **knows** the facts. (The Times)
2. You're one of the few who **was** aware of their existence. (BNC)
3. Nusseibeh is one of the few who enthusiastically **support** the deal. (Time)
4. Heaney is one of the few poets who **loves, honours** and **cherishes** himself and his ancestors in his poetry. (The Times)
5. He is one of the few people I have ever met who **has** never

been either inflated or deflated by personal possessions. (BNC)

6. He is one of the few middle-aged politicians who **look** more virile in a swimsuit than in a business suit. (Time)

A closer analysis of examples in the three corpora under investigation shows that, in short, the plural form is more frequent than the singular. In the BNC, 21 out of 38 examples are plural (55%). There are notable differences in the other collections of data. Of the 47 examples which are in the structural pattern [one of the few + noun + (who) + verb] in *The Times*, 33 examples are plural (70%). Of the 10 examples in *Time*, 8 examples are plural (80%). The plural form occurs more frequently in these two corpora than in the BNC.

From this observation we can come to the conclusion that it is more common for the plural form of the verb to be used (in total, 62 out of the 95 examples above), but the preference is not overwhelming. The singular form is used sufficiently enough to be regarded as an acceptable alternative.

Logically, the relative clause defines the group from which an individual is singled out, and plural concord would seem to be the natural choice (as in 1, 2, 4, 5). It is also the more frequent choice. The cases of singular concord (as in 4) should probably be ascribed to the pull of the numeral *one* towards the singular, combined with the fact that the main clause makes reference to a single person.

3.2. Concord with Measure Expressions

We will now turn to the concord with measure expressions, with special reference to the measure word *inch*. Which verb form should the following sentence take? Singular or plural.

Up to 10 inches of rain **has/have** fallen in some areas since Sunday.

The question here is whether to treat the subject in the sentence

as singular or plural. According to Biber *et al.* (1999, p. 187), however, the general rule is that plural measure expressions (e.g., amount, weight, length, time, etc.) take singular verb forms, if the reference is to a single measure.[6]

Let us first look at the following examples taken from the BNC:

1. There **is** about 15 inches of snow in the village, with drifts and white-outs making driving hazardous.
2. Up to three inches of snow **is** expected to fall over eastern and some southern areas of England today, the London Weather Centre forecast.
3. In Scotland up to three inches of snow **was** hampering drivers on many roads in Dumfries and Galloway, although most routes were open by yesterday morning.
4. ...where hundreds of inches of rain **fall** every year and spectacular waterfalls cascade through rocky canyons on to deserted black-sand shores.
5. More than eight inches of rain **has** already fallen in Somerset. In Wales, hundreds of acres of farmland around Welshpool **were** under
6. 17 inches of rain **have** fallen in the mountains above Santa barbara so far this year.
7. Weather forecasters said several inches of snow **were** likely in places.
8. Six inches of snow **were** reported in the resort of Scarborough.
9. But he might prove a better bet than either at international level where inches of space **are** precious and the chances to storm past defenders rare.
10. We calculated that about 20 inches of water **were** needed to cause capsize.

We can find examples of number concord in which the subject (including the measure word *inches*) is treated as both singular and plural. As these examples show, English speakers seem to be uncertain about the rules of the concord with measure

---

[6]Biber *et al.* (1999, p. 188) gives the following examples:
[Two pounds] is actually quite a lot. (CONV)
[Eighteen years] is a long time in the life of a motor car. (NEWS)

expressions. Quirk *et al.* (1985, p. 757) states that difficulties over concord arise through occasional conflict between grammatical concord, notional concord and the principle of proximity.

The principle of notional concord accounts for the common use of a singular with subjects that are plural noun phrases of quantity or measure. The entity expressed by the noun phrase is viewed as a single unit, as the following two examples show: *Ten dollars is all I have left; Two thirds of the area is under water.* On the other hand, Biber *et al.* (1999, p. 189) notes that the regular pattern of grammatical concord may be disturbed by the principle of proximity, as the following example shows: *One of the girls have got bronchitis.*

In the case of the measure word *inch*, the examples taken from the BNC show that the plural is the preferred choice. It should be noted here that any conclusive answers to questions about grammar cannot easily be drawn from traditional studies of language structure based on intuition, anecdotal evidence, linguist's knowledge of language. Rather, they require empirical analysis of large databases of authentic texts.

4. (Reflexive) Pronoun Choice in Coordinated Noun Phrases

Two or more noun phrases may be conjoined to form a coordinated noun phrase, which can occur in the sentence as subject, object, or complement and may have a noun or pronoun as their head. In this section we will look specifically at the coordinated noun phrase [(pro)noun + x-self] occurring as subject.

(Reflexive) pronoun choice in coordinated noun phrases could be used to test the extent of a learner's knowledge of English grammar. In a multiple choice format a test writer can create a question like the following:

A: What are you going to do tomorrow afternoon?
B: My brother and _____ are going to the movies.
(a) I (b) mine (c) me (d) myself

The question here is which form of the pronoun is the correct answer. It has been widely accepted that in addition to different order preferences,[7] the pronouns have different case preferences in coordinated noun phrases. This can be seen in the following example sentences taken from the BNC:

1. You and **I** have something in common you know.
2. Nicole and **I** are getting married.
3. Annabel and **I** are supposed to help distribute the food.
4. Tony and **me** have come down here again.
5. Bella and **me** both saw him.
6. Why don't you and **me** go some place?
7. No matter how much we are in love, when you and **me** look out of the same window, we do not see the same things.

In principle, the nominative form is used in subject position, while the accusative form is used in object position. In the case of coordinated noun phrases, however, accusative pronoun forms are more frequently used in subject position, especially in conversation. In fact, the general tendency is for the accusative pronoun forms to spread into contexts traditionally associated with the nominative case.

Note here that, like the accusative pronoun forms, reflexive pronouns can also be used in coordinated noun phrases along similar lines. Biber *et al.* (1999, p. 339) points out that reflexive pronouns are occasionally found in coordinated noun phrases in subject position, as shown in the following examples:

---

[7]According to Greenbaum & Quirk (1990, p. 274), it is considered polite to follow the order within a conjoint noun phrase of placing 2nd person pronouns first, and (more importantly) 1st person pronouns last: *Jill and I; you and Jill; you, Jill and me.*

Paul and **myself** went up there didn't we? (CONV)
"My three associates and **myself** are willing to put big money into the club to get the best players for the team." (NEWS)
Only **myself and my family** are affected by it. (NEWS)

Biber *et al.* (1999, p. 339) contend that reflexive pronouns, which have no case contrast, provide a convenient way of avoiding a choice between a nominative and an accusative case form. They further comment that examples of this kind are generally rare and occur mainly in news.

Now let us look at more corpus data to see how much more information we can obtain than the grammar book can supply. Consider the following examples taken from the BNC:

1. Rose and **myself** are away for the day.
2. So Steve and **myself** are er have done a done a bid that we should know...
3. Both Chris Armstrong and **myself** are fairly new in the team...
4. Who's calling? Sergeant, it's PC Garfield, Sergeant Huddersfield and **myself** are at the Riverside Hotel...
5. then Neil Henshaw and **myself** climbed aboard.
6. Not only was there Mikey and **myself** but in the year above me was Gary Fraser.
7. Tony and **myself** come on and we do an hour and a half...
8. I think the both Mr and **myself** considered that it might er assist your Lordship...
9. I was wondering if you, Margaret and **myself**, could get together soon to discuss them?
10. He had not. In the meantime, Timpson and **myself** had planned to strafe the road...
11. Fortunately, Ewen and **myself** had remembered to bring along the Trophy...
12. The cast was young, and only the producer and **myself** had seen any of those days.
13. One of the aims of both Tim Grant and **myself** has been to build up expertise...

It seems true that the use of reflexive pronouns can avoid the choice between a nominative and an accusative case form. There

seems to be another reason for choosing the reflexive pronoun. Speakers are often said to resort to reflexive pronouns for emphasis. The use of emphatic reflexive pronouns can be regarded as an example of this tendency.

Now going back to the grammar question given earlier in this section, some test writers seem to think that examples of this kind are considered to be generally rare, impossible, or grammatically incorrect. It is also very likely that test writers who advocate prescriptive grammar choose to be 'prescriptive, not descriptive'. As these examples show, however, corpus-based analysis provides ample evidence that the coordinated noun phrase [(pro)noun + myself] is perfectly acceptable as subject, object and complement, although the use of reflexive pronouns are less frequent than the use of personal pronouns.[8]

What seems particularly noteworthy here is that these examples occur primarily in spoken English, while being relatively rare in written English. This, however, contrasts with the statement of Biber *et al.* (1999, p. 339) that examples of this pattern occur mainly in news.

One of the most important uses of corpus-based investigation is to provide information about frequency of occurrence in real language use. Hitherto this information has been based on native-speakers' intuition or test writers' knowledge about language. Test writers will also find corpus-driven information useful to learn which form or structure is common and which are rare or impossible.

5. Conclusion

[8]Because of the cumbersome process involved in extracting information about the frequency of each pronoun form in the structural pattern under consideration from the corpus, at the moment it is difficult to compare the distribution of each pronoun form in statistical terms.

In this paper, particular lexical and grammatical features of English were explored from a corpus-based perspective. Some of the ways in which corpus-based investigations can shed light on the study of English grammar have been demonstrated. A corpus-based approach attempts to uncover typical patterns rather than making judgments of grammaticality based on speakers' intuitions. This approach has made it possible to conduct statistical investigations into the frequency of grammatical patterns in use and thus to expand the scope of earlier investigations. It can therefore bring to the fore many aspects of language structure and use that have not received attention in traditional studies.

Native speakers rarely have accurate information about frequency of use. A large representative corpus is the only reliable source of information about frequency and use. It is therefore important in many ways to base one's analysis of language on real data rather than data that are contrived. Any description of grammar would be more complete and accurate if it were based on a body of real data.

Before the advent of electronic corpus data, the quantity of data was inadequate for any reliable statements about vocabulary, grammar and usage. Computers have recently made it possible to carry out the analysis of a large quantity of natural language and identify patterns of language use. Although evidence from very large corpora sometimes causes a conflict between introspection and real language use, the last decade has witnessed (and this decade is still witnessing) a growing need for corpus evidence.

References

Aijmer K. and B. Altenberg. 1991. *English Corpus Linguistics.* Harlow: Longman.

Aston, G. and L. Burnard. 1998. *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, D. *et al.* 1999. *Longman Grammar of Spoken and Written English.* Essex: Pearson Education Ltd.

Burnard, L., ed. 1995. *Users Reference Guide for the British National Corpus,* Oxford: OUP Computing Services.

Greenbaum, S. and R. Quirk. 1990. *A Student's Grammar of the English Language.* Harlow: Longman.

Jackson, H. 1988. *Words and Their Meaning.* Harlow: Longman.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* London: Longman.

Landau, S. 2000. *Cambridge Dictionary of American English.* Cambridge: Cambridge University Press.

Lewis, M., ed. 2001. *Teaching Collocation: Further Developments in the Lexical Approach.* Hove: Language Teaching Publications.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McNamara, T. 2000. *Language Testing.* Oxford: Oxford University Press.

Meyer, C. 2002. *English Corpus Linguistics.* Cambridge: Cambridge University Press.

Quirk, R. and S. Greenbaum. 1990. *A Student's Grammar of the English Language.* Harlow: Longman.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvic. 1985. *A Comprehensive Grammar of Contemporary English.* London: Longman.

Rundell, M. 2002. *Macmillan English Dictionary.* Oxford: Macmillan Education.

Simpson, R. and J. Swales, eds. 2001. *Corpus Linguistics in North America.* Ann Arbor: The University of Michigan Press.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. McH. 2001. *Collins COBUILD English Dictionary for Advanced Learners.* Glasgow: HarperCollins.

Summers, D. 2000. *Longman Advanced American Dictionary.* Harlow: Pearson Education Limited.

Swan, M. 1995. *Practical English Usage,* Oxford: Oxford University Press.

Thomas, J. and M. Short, eds. 1996. *Using Corpora for Language Research.* Harlow: Longman.

Wichmann, A., S. Fliegelstone, T. McEnery, and G. Knowles. 1997. *Teaching and Language Corpora.* Harlow: Addison Wesley

Longman.

Heok-Seung Kwon
Department of English Language & Literature
Seoul National University
San 56-1, Shinrim-dong, Kwanak-gu, Seoul
151-742
Phone: 02) 880-6357
E-mail: hskwon@snu.ac.kr

received: January 5, 2004
accepted: February 15, 2004