The use of language corpora in the teaching of English

Tony T.N. Hung
Hong Kong Baptist University

Abstract

While corpus linguistics has firmly established itself as an essential tool for linguistic research, it has yet to win converts from the ranks of language teachers as an invaluable resource in language teaching.

Specially written for teachers of English with little background or interest in linguistics, the paper is an introduction to the ways in which English corpora (such as the Bank of English) can be used in checking and updating the teachers’ own understanding of the state of the language as it is actually used today, in designing language-learning tasks, and in raising the learners’ consciousness to the differences between their interlanguage grammars and standard English grammar.

Introduction

Corpus Linguistics is a branch of language study which engages in the systematic and extensive collection of real language data, both written and spoken, as a basis for linguistic analysis and description. Advances in computer technology in recent decades have made it possible for massive ‘corpora’ or ‘databanks’ of hundreds of millions of words to be collected, and for the data to be tagged and categorised, and retrieved with ease for various purposes.

Language corpora have become such an important tool for linguistic research, and especially for lexicography, that nowadays it has become almost unthinkable to write a dictionary or a descriptive grammar of a language which is not (at least to a large extent) corpus-based.¹ Given the nature of language as a means of communication, the importance of authentic data from language use is perhaps too evident to require comment. Biber et al. (1998) provide a

helpful introduction to the various uses of corpus linguistics for language research – in lexicography, grammar, discourse analysis, register variation and English for Specific Purposes.

The increasing importance of corpus linguistics for practitioners of language research is, however, not paralleled among practitioners of language teaching. In fact, most language teachers today are not even aware of what language corpora can do, let alone make use of them in their teaching. This lack of interest stems perhaps from a broader reluctance among language teachers to familiarise themselves with, and make use of, the findings of modern linguistics in general in their teaching (a theme addressed by Hung 2001). In the present paper, I shall attempt to show how even language teachers with relatively little background in linguistics can benefit from the use of language corpora in the teaching of English. I shall not refer to its usefulness for linguistic research as such (of the kind mentioned by Biber et al. above), but only for language-teaching, particularly the teaching of grammar.²

The facts of language

Before anyone can teach a language, they have to know at least the ‘facts’ – i.e. the state of the language as it is actually used today. But where are the ‘facts’ of grammar – as opposed to the theories – to be found? There was a time when self-appointed authorities (like the 18th century grammarian Robert Lowth) used to lay down the law on what was grammatical or ungrammatical, even when it went against common usage or common sense. Lowth’s most famous saying was that ‘It is not the language but the practice that is at fault’ – in other words, the language itself was perfect but the users were not. We have come a long way since then. One of the most important lessons of modern linguistics in the last century was that the grammar of a language resides not in grammar books but in the practice of its speakers, who are the ultimate ‘authority’ for the current state of the language.

The importance of language corpora for teachers is that they provide readily accessible, comprehensive and up-to-date sources of data on the current state of the language, in all of its aspects. Through them, teachers will

often discover that some long cherished prescriptive notions of grammar or usage are not borne out by the language as it is actually used by educated people today.

Take for example the use of interrogative who and whom. Most prescriptive grammars would tell you to use whom when the pronoun functions as the object in the clause, e.g. ‘Whom did you see?’, ‘Whom did you invite to the party?’. But for decades now, increasing numbers of users of English have defiantly stuck to who (‘Who did you see?’, ‘Who did you invite to the party?’). The trend alarmed even someone like H.W. Fowler, who was normally the most level-headed of prescriptivists, as evidenced by the tone and choice of words in this entry from his Modern English Usage (first published in 1926):

This colloquialism [who] is indeed so common that it is invading printed matter... But who’s invasion of the province of whom has not gone so far in..., and we may reasonably suspect such sentences...to be due to carelessness rather than a splendid defiance of grammar.’ [emphases added]

‘A splendid defiance of grammar’ – as if grammar had an independent existence from, and authority over, the users of the language. We have no statistics on the relative frequency of who and whom in Fowler’s time, but now we do. Here is a table from the Longman Grammar of Spoken and Written English (1999), which is based on an extensive corpus of spoken and written data:

Table 1: Interrogative who and whom
[No. of tokens per million words in the Longman Corpus]

<table>
<thead>
<tr>
<th></th>
<th>who</th>
<th>whom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversation</td>
<td>1,000+</td>
<td>0</td>
</tr>
<tr>
<td>Fiction</td>
<td>350+</td>
<td>&lt;25</td>
</tr>
<tr>
<td>Newspapers</td>
<td>150+</td>
<td>&lt;25</td>
</tr>
<tr>
<td>Academic writing</td>
<td>50+</td>
<td>&lt;25</td>
</tr>
</tbody>
</table>

[from Biber et al. 1999: 214]

These figures show that the use of interrogative whom has virtually disappeared from conversational English, and that its occurrence in written English is very negligible and far below that of who, even in academic writing.
The use of language corpora in the teaching of English

It would not be too far-fetched to predict that it will disappear completely from English usage in the not too distant future.

Language corpora as a window of modern usage

A corpus-based grammar like the Longman Grammar is a convenient way of looking up facts such as the above. But for the most comprehensive and systematic access to modern-day language data, teachers (or at least their schools) can subscribe to one or more of a number of large corpora available via the Internet today, such as the Bank of English (Collins Cobuild), British National Corpus, International Corpus of English, Brown Corpus (for American English), Wellington Corpus (for New Zealand English), and so on [see the References section for their respective websites]. The Bank of English corpus, for instance, consists of about 300 million words, and is constantly being expanded and updated; the data are taken from written and spoken sources over the last 10 years, from Britain, America and Australia, and are therefore up-to-date and international in character.

In the rest of this section, I shall provide a number of examples from the Bank of English to illustrate what teachers can learn from language corpora about the English language today, which they might not from a conventional textbook. But first, a general observation. One of the most striking things that one notices when engaging with real language data is the fact that there is much more variation in a language like English than most teachers would care to admit or allow. I am not referring to learners’ errors, but to ‘standard’ English as produced by educated users, and appearing in newspapers, books, broadcasts, etc. Teachers, unfortunately, have the reputation of being intolerant of variation: to them, there can only be one correct form for everything and one correct answer to every question. And if you tell them that there are two (or more) forms which are both acceptable, they think that you are unsure of the ‘facts’ of the language, when in fact it is quite the opposite. This is perhaps a by-product of an examination-oriented mentality, but it is a distortion of the true nature of language nonetheless.

At the simplest and lowest level, language corpora allow us to check on the relative distribution of morphological variants in the English of today. Take the plural form of the word syllabus, for which two variants are given in most dictionaries, i.e. syllabi and syllabuses. Which is the preferred form
among English users today? A quick check on the occurrences of both forms in the Bank of English shows the following number of tokens:

Example 1: syllabuses vs. syllabi 
[No. of tokens in Bank of English]

<table>
<thead>
<tr>
<th>syllabuses</th>
<th>syllabi</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>1</td>
</tr>
</tbody>
</table>

Evidently the form *syllabuses* is heavily favoured by present-day users over *syllabi*. Given this fact, and the fact that *syllabuses* is regular and therefore much easier to learn, there would seem to be no logical reason to teach students the form *syllabi*.

Moving on to a syntactic structure, let us take another irregular form, *if I were*, which is one of the very few surviving forms of what used to be called the ‘subjunctive’ in English. What about the regular form, *if I was*? One finds the following number of tokens for each form in the Bank of English:

Example 2: if I were vs. if I was 
[No. of tokens in Bank of English]

<table>
<thead>
<tr>
<th>if I were</th>
<th>if I was</th>
</tr>
</thead>
<tbody>
<tr>
<td>345</td>
<td>647</td>
</tr>
</tbody>
</table>

A cursory examination of the data shows that the great majority of the *if I was* tokens (over 75%) are used in the ‘subjunctive’, as opposed to the simple past tense use. After making those allowances, the data still show that *if I was* is now more frequently used than *if I were* in the subjunctive sense, and though the former may be more common in speech, the data come from both spoken and written English. I am not saying that teachers should not teach *if I were*, but if our students use *if I was*, I do not think they should be faulted, or else over half of English-users today should also be faulted.

In most traditional grammars, the quantifier *less* is supposed to apply to singular mass nouns, as in *less money*, *less time*, and *fewer* to plural count nouns, as in *fewer people*, *fewer books*. It is interesting to note, however, that *less* has begun to gain currency with plural count nouns as well as singular mass nouns. The Bank of English gives no less than 153 tokens of this use, e.g.:
Example 3: *less* + plural noun

[The following are some examples from a total of 153 tokens in Bank of English]

is also worried. He says there are *less teachers* to choose from when jobs in the trees scales down - because *less leaves* are being eaten and broken - presumably because there were *less parasites* around to be transmitted a media of today where they're taking *less chances* and basically every film is back into the state in return for *less expenditures* in attempting to try to specifics, I think you find *less and less people* who will support things. Yes, I did expect more food and *less people*. That I expected would be Recipient. It's better because you get *less ways* of getting stuck up or spending less and therefore paying *less taxes* on income and property. So has to give. This translates into *less benefits* or higher premiums. jobs. And these companies with 20 or *less employees* typically are paying low Fewer people working, you've got *less taxes*. Fewer people working, you smaller pieces and therefore they had *less calories* and fat, when in fact, the What I want to do is give them *less taxes*, less regulation and less state system. This would mean even *less resources* available to schools. afraid that later Koen will have *less opportunities* in life because of his to the consultants, would produce *less accidents* because fewer pedestrians scrip was weak, J.B Were could hold *less shares* back for its own clients. by advisers which will result in *less funds* flowing into BT in the short Mr Fitzgerald said. Less and *less people* are taking them up, and I to supervision. It results in *less people* in prisons," he said. All *s work. Well, there's less and *less people* in the country. In the usually get the same model with two *less doors*. We are offering them a real country music's Mike Flowers, *less mavericks* and more mimics. They easier to assimilate and would cause *less problems* for the people of Virginia Everything's just the same except *less people* are going to watch it. By small businesses, they might have had *less profits*, but their customers would for next season. There will be *less clubs* in the Premier League by then Lomu, 20, would probably have had *less games* than Matt Le Tissier has played singly. In that way we would risk *less casualties* if they have grenades. your whole attention and to give them *less results* in errors or

While it is true that about 400 tokens of *fewer* + plural nouns are also given in the corpus, the point remains that *less* is no longer ‘sub-standard’, but is gaining currency among educated users today. Personally, I have no doubt that it will displace *fewer* completely in the not too distant future, if for no other reason than that it is both simpler and more symmetrical, as the opposite of *less* is *more*, which applies equally to both kinds of nouns: *more people*, *more money*.
Mathematical logic is often invoked to settle arguments over grammar and usage – for example, a sentence like ‘I didn’t see nobody’ is said to mean ‘I saw somebody’ because (mathematically) two negatives make a positive, in contrast to its obvious intended meaning where the double negation serves to reinforce the negative meaning. But natural human language is not mathematics. A good example of how linguistic usage overrides mathematical logic is the use of the singular pronoun everyone or everybody with a plural pronoun. This practice has become increasingly common over the past twenty years or so, as a strategy for avoiding the use of the sexist pronouns he, his, himself to refer to everyone/everybody, or the unbearably clumsy he or she, his or her, himself or herself. To more and more people, the use of the plural they, their, themselves seems like the most expedient and sensible solution to the problem. The following data consist of the first 20 entries from the Bank of English involving everyone followed by a pronoun. It shows that plural pronouns are preferred to singular in the majority of cases involving everyone. It is no use invoking mathematical logic (as some teachers do) and saying that there is a mismatch between singular and plural. The users of the language have declared their preferences.

Example 4: everyone + singular or plural pronoun
[first 20 entries from Bank of English]

<table>
<thead>
<tr>
<th>first 20 entries from Bank of English</th>
</tr>
</thead>
</table>
| take action. Ray suggests that everyone must ask themselves one central in Scotland; it offers a chance for everyone to pull themselves together after versus the British of course, everyone is out for himself in this game of told her it would be a while. 'Is everyone behaving herself? And him?' 'It' The navy, he added, should destroy everyone who revealed himself as a traitor. he discovered, was a business. Everyone was out to enrich himself, and was worthy of all that bullshit that everyone puts themselves through in order are entering the country, and that everyone in Yugoslavia is arming himself as but the other did not. For everyone who exalts himself shall be fun for all. Not only did everyone enjoy themselves, but over $dollar; hand round the sauce in a bowl for everyone to help themselves. Rice and new vinyl costs a lot more-why doesn’t everyone go out and buy themselves a cheap creating more hurt than good. Everyone must think for themselves, not against some Mickey Mouse teams and everyone is falling over themselves to pay backer giving half that amount, everyone would be falling over themselves The worst thing has been watching everyone training and enjoying themselves I thought we’d get a draw. Now everyone must pick themselves up.” That is much. She had little ways of making everyone feel better about themselves. If since I have been at the club. Everyone is enjoying themselves - including Arthur makes the arrangements and everyone enjoys themselves. Advocaat is a
Still on the subject of number, a good deal of argument has been wasted on things like whether the pronoun none is singular or plural. The fact is that users of English treat it as both, and there do not seem to be any identifiable variables that condition the choice of one or the other, as evidenced by the following entries from the Bank of English.

Example 5.1: none + singular verb

with a remarkably even-handed film. None of his characters is truly heinous—would be highly improbable and none of the crew is implicated, says the go home. At their own request, none of the girls is being identified. But of supermodels like Naomi Campbell – none of the stars is black. The new soap write my copy and make sure that none of my colleagues is gullible enough British composer Judith Weir. None of our composers is as adept as A decade later, of course, none of our critics is volunteering to business in the next few years. None of these companies is wholly focused On the present showing, none of those conditions is likely to be the park itself is small and none of the buildings is unduly grand. about 40 of the time. Likewise, none of these securities is subject to known words, and cross-check meaning. None of these activities is prerequisite to be built at the airport. None of the streets is named on the map say, Ravel’s Bolero or Nutcracker but none of the composers is likely to be

Example 5.2: none + plural verb

smuggling the food in on mules, and none of the men are starving. Despite Prof. WILSON: Well, at the moment, none of these organizations are doing has been affected by a virus. None of her runners are likely to be through their individuality. None of mine are like that. We've only lay clients with a service second to none and its members are justifiably in exactly the same way. None of my friends are girlie types abnormal menstrual bleeding. However, none of these symptoms are very specific I'm beginning to be afraid, because none of my feelings are coming back. One a lover, a hunter, a fighter, and none of those instincts are given much away who are in need," she said. None of the proposals are definite just some of whom would be Australians, none of whom are responsible for the genre, isn't it strange that none of the above are fashions for men? jobs unpaid. Clach made it clear none of their players are guilty of with closure due to lack of cash. None of the staff are paid so the
be entitled to top billing. But none of the teams are likely to be
But they're not all on a manhunt. None of the girls are married, and only
Teachers who are aware of such facts of usage should be open-minded enough
to accept both from their students, rather than arbitrarily insisting on one or
the other.

A final example comes from another area of grammar which has been
undergoing change. In traditional grammars, the verb form which follows
such verbs as recommend and suggest is supposed to be in the ‘subjunctive’
(e.g. ‘I recommend/suggest that she return the money immediately’), which in
this case is identical to the non-finite, uninflected form. However, the regular
inflected form is now gaining acceptance, such that in the Bank of English
there are just as many tokens of that as of the subjunctive form. For example:

Example 6.1: recommend that + finite verb form
[entries from Bank of English]

In any case Mr Chapman would not recommend that Charles puts away his polo
she receives substantial bonuses, I recommend that she pays pound 150 a month
at Greater Manchester, which will recommend that Britain does not set up an
of Defence is today expected to recommend that Britain joins France and
non-binding resolution which would recommend that the President does not
Mr Javier Perez de Cuellar, is to recommend that the UN adopts sweeping new
Mr Javier Perez de Cuellar, is to recommend that the organisation adopts
marches and Markmen's parades. We'd recommend that everyone visits the flea

Example 6.2: recommend that + non-finite verb form (‘subjunctive’)

500 billion target, he would recommend that the president reject the
Yitzhak Shamir has said he will recommend that his country participate in
attacks, we're not in a position to recommend that anybody take anything.
Wren says he, too, isn't ready to recommend that everyone take vitamin E
and asked me whether or not I would recommend that she start using an anti-
area as well as other regions. We recommend that the woman simply incorporate
particular family. Some theorists recommend that the therapist assess
case, urged Justice Ambrose to recommend that the Crown pay all parties'

Now, as individual users of the language themselves, teachers are of
course as entitled as anyone else to their own preferences in the use of
language. But as teachers, it is their job to teach the language as it is actually
used today, not someone’s preferred version of the language. Corpus linguistics and modern surveys of grammatical usage are an essential source of information which helps teachers to decide on what is current and what is out-dated in teaching the grammar and lexis of the language.

**Language corpora as a resource for designing language-learning tasks**

The main contribution of corpus linguistics to language teaching (as opposed to linguistic research) is thus to provide the teacher with crucial evidence for the state of the language as it is used today. But corpus linguistics can also serve as an almost inexhaustible resource for language teachers in constructing language-learning tasks. The readily available data not only save teachers the time and trouble of making up their own sentences and texts, but have the important advantage of being authentic and up-to-date examples of language use.

There are various ways in which corpus data can be adapted for use in the classroom. For example, data on the use of a particularly difficult or unusual grammatical construction (such as ‘it is time that…’) can first be presented to learners in such a way as to help them notice the special features of the construction. A small sample of data such as the following would suffice for purposes of getting learners to come up with their own generalisation about the linguistic features entailed by this construction:

**Example 7: it is time that...**

solution. Jaffe: I think it is time that voters **took** responsibility for school books. It is time that books taught at school **were** a minimum income in old age”. It is time that Mr Smith **apologised** for structure of the European game. It is time that it **came** down from the clouds. sub-standard performance. It is time that comfort zone **was** removed and be in a tight spot. But maybe it is time that you **were** forced into a position of technical change, and it is time that sociologists again **attended** to tend to be too sweet, then it is time that you **discovered** just how good for good reading, but maybe it is time that one airline **learned** its lesson

Most learners will probably arrive on their own at the generalisation that the construction ‘it is time that…’ entails the use of the past tense form of the verb in the dependent clause that follows. In this way, they can learn to discover grammatical rules inductively, rather than being told explicitly and without
exercising any mental effort. As a follow-up to such tasks, learners can be given further sentences from the corpus where they are asked to fill in the blank with the correct form of the verb given in brackets:

**Exercise**

Fill in the blanks with the correct form of the verbs in brackets:

a. I think it is time that voters _____ (take) responsibility for their choices.

b. It is time that Mr Smith _____ (apologise) for his behaviour.

c. It is time that you _____ (discover) just how good you really are.

d. Maybe it is time that airlines _____ (learn) their lesson.

**Learners’ corpora**

Learners’ English, as most teachers will have noticed, differs from standard English in systematic ways (cf. also the many studies on learners’ ‘Interlanguage’, e.g. Selinker 1972, Yip 1995). This is apparent even from small samples of learners’ English (such as the following, from Hung 2000):

1. A large majority thought the presenting skill was important.
2. There are different opinions about leadership ability, writing ability and selling ability are important or not.
3. More than 50% of sample disagree attractive appearance is necessary for...
4. Respondents considered the attractive appearance was not an important quality.
5. Africa countries adapt super carrot and find the super carrot can inhibit...

As a ‘consciousness-raising’ exercise, teachers can ask the learners first to examine the above data from their own writings, and to compare them with the following data on similar constructions from the Bank of English, and to analyse their differences:

1) **think (thought) that**...

   Well, in this case, the experts thought that Oakland might sweep this
   Over 50 percent thought that he should resign. Seventy-
   Seventy-some percent thought that he should not run for
   Some people thought that the church was on fire.
Other people thought that there had been another murder laid off in September, but I never thought that I would be laid off this. Connely: I thought that was very significant. I they must have thought that they didn't have a history, and so we, of course, thought that it was the silliest thing in

2) consider that...
And when you consider that inflation here is running at especially when you consider that it means overturning hundreds. She didn't really, you know, consider that we were just kids. But now, and they didn't consider that the economic problems were well, you see, I consider that everything I do is a form of which is a serious mistake when you consider that men traditionally rule the Mining the Museum asks viewers to consider that there are several ways at and Kidney Diseases. If you consider that the body is made up of all you, says that he is Hungarian, I consider that he is not loyal to the improbable, especially when you consider that most cats do not like each

3) opinion...whether...
may be right in his nautical opinion about whether it was wise to bring Cowley declined to give an opinion on whether Opposition Leader John children's viewing hours but had no opinion on whether there ought to be more and were more likely to have an opinion on whether hand washing was I do not have an opinion on whether George Michael should Ferguson is entitled to his opinion whether Hoddle regards those views Normally an agency does not give an opinion on whether you are a suitable Should doctors revise their opinions about whether patients diagnosed There were also radically different opinions on whether NATO should adopt I'm sort of I've got mixed opinions about whether it's best to direct

Such comparisons and analyses may help learners to notice how standard English grammar differs from their own interlanguage grammars with respect to this particular construction.

One final example of the use of learners’ corpora and standard language corpora comes from the existential construction (there is/are...). Many Hong Kong students produce sentences such as the following:

1. There are over 80% of them agreed with that.
2. There are about 23 percent of them disagreed with the importance of leadership ability.
3. There are students study in the library.
4. There were more and more competitors entered the market.
5. There are many people travel abroad.

As analysed in Hung (2000), such sentences apparently stem from an interlanguage rule that allows one simply to attach ‘there is/are’ to the beginning of a regular sentence, such as ‘Many people travel abroad’ and ‘More and more people entered the market’. The results are (of course) ungrammatical in standard English. Students can be given the task of comparing the above ungrammatical sentences with the same construction from a standard English corpus, and analysing their differences:

4) *There is/are...*

there are children who are coming to school in a great
Indeed there are hardliners who are by no means satisfied that
There are people who will not allow people from projects
outside Iraq, there are populations who support him—in Jordan, among the
protagonist tells us there are many who would have considered it a blessing
There are many who would say that the mayor at the same
I think there are women who would feel uncomfortable if they're
There are some who believe that it will and that growth
And there are some who argue that it should be legal
And there are people who will argue that's an important
But there are others who argue that, like the conflict
to stamp it out, but there are others who prefer that the old system
Of course, there are many who disagree with the mayor, Mikhail
and then there are people who just commit a crime. And Soon Ja

Finally, it would be necessary to remind ourselves that, however useful it may be, corpus linguistics is only a tool, and that ultimately it is the teacher’s own linguistic sense, knowledge and experience which will determine how wisely and productively this tool is used in the teaching of a language.
References

Bank of English website: www.cobuild.collins.co.uk/boe_info.html


British National Corpus website: http://www.hcu.ox.ac.uk/BNC/index.html

Brown Corpus website: http://www.hd.uib.no/icame/brown/bcm.html


International Corpus of English website: http://www.ucl.ac.uk/english-usage/ice/


Wellington Corpus website:  
http://www.vuw.ac.nz/lals/wgtn_crps_spkn_NZE.htm