Antoinette Renouf

The Establishment and Use of Text Corpora at Birmingham University

Abstract

The School of English at Birmingham University has over the last ten years increasingly integrated the study and use of corpora into its research and teaching activities. Cobuild Ltd and the English for Overseas Students Unit are particularly active, as is the Research and Development Unit for English Language Studies. Members of the Research Unit have created the purpose-built corpora that make up the Birmingham Collection of English Text. The Research Unit is using these to support its linguistic research projects and the development of new types of text-processing software, as well as for specialised teaching purposes.

1. Introduction

There is a wide range of innovative corpus-based work taking place within the School of English at Birmingham University, most of which is best reported on by the people concerned. I shall therefore only touch briefly on some of the latest work in two areas, Cobuild Ltd. and EOSU (English for Overseas Students Unit), before focussing on the work of my own unit, the Research and Development Unit for English Language Studies.

1.1 Cobuild Ltd

The Cobuild project has, over the years, produced a large number of reference and teaching works that are most probably known to the reader, the first being the ‘Collins Cobuild English Language Dictionary’ and the most recent the ‘Collins Cobuild Student’s Grammar’, and the first two ‘Collins Cobuild English Guides’, to prepositions and word formation respectively. The evolution of this pioneering project, from its beginnings in 1980 until 1987, has been described in detail in Sinclair (ed), 1987.

The basic data resource for the project has hitherto been the 20 million word Birmingham Corpus of writing and speech. But Cobuild Ltd. is now embarking on a series of major new initiatives, one of which is the creation of ‘The Bank of English’, a new corpus of speech and writing containing hundreds of millions of words. It is felt time to create a resource that is both up-to-date and large enough to reveal information even about the hitherto elusive rarer words. The corpus will be assembled fairly speedily, taking data that is as far as possible already available in electronic format, such as published books, newspaper data and radio programme material, although work on recording the spoken word is also under way. This new corpus initiative was recently announced to the general public, and was widely reported in the British press.

1.2 EOSU

In the English for Overseas Unit, Tim Johns has for some years been developing what he calls ‘data-driven language learning’ (Johns, 1989). This is a novel way of allowing students to be active learners of English, by getting them to carry out their own study of raw data from specialised corpora. Students use a range of microcomputers to extract concordances, word lists and so on from the text, thereby releasing the teacher from the burden of being the sole language informant in the learning process, and employing him or her as co-researcher instead. Tim’s own students have taken very well to this approach and have made important discoveries about English grammar and vocabulary that, Johns has recently reported (1990), have ‘left no escape from the conclusion that the description of English underlying our teaching … needs radical reassessment’.

2. The Research and Develpoment Unit for English Language Studies

The Research and Development Unit grew out of the original Cobuild project, and was set up in 1985 as a self-financing entity within the School of English. Its chief purpose is to carry out fundamental corpus-based linguistic research, primarily in the English language, with a particular focus on lexis and collocation. This inevitably involves the creation, from time to time, of new data resources and tools that are adequate to support the innovatory research goals. The Unit’s main areas of activity have so far been:

- creating new, predominantly English, text corpora of all kinds
- carrying out corpus-based linguistic research
- developing new text-processing software
- cooperating in School and outside corpus-based projects
- providing corpus data for internal and external users
- corpus-based teaching

2.1 Corpus Creation

2.1.1 The Birmingham Collection of English Text

Over the last ten years, members of what is now the Research and Development Unit have created the purpose-built corpora that reside under the umbrella of the Birmingham Collection of English Text. These include the afore-mentioned 20 million-word Birmingham Corpus, the one million-word TEFL Corpus, and task-based spoken corpora. The process of design and construction in these cases has been reported on at various stages (Renouf, 1986, 1987). The Collection also contains large amounts of data acquired from individual sources, such as a 13 million-word corpus of speech from a public enquiry, and data from 'The New Scientist', 'Byte' and 'Nature' journals, and from the Times and other newspapers.

2.1.1.1 The 'New Corpus of Spoken English'

Among the Research Unit's on-going projects is the creation of the 'New Corpus of Spoken English', that began four years ago as a background activity. About one million words of speech has already been recorded, and half of this transcribed, but corpus builders will know how very slow the process is. Even orthographic transcription of this amount of text is a large task, and prosodic coding, which we are not undertaking, would make the task exponentially greater still.

The corpus has yet to find its final shape. Since we are building it piecemeal, we have time to mould it in accordance with our growing understanding of corpus design. Recently, we have been concentrating on building in a large component of undergraduate informal conversation. This data is a rich source of studentisms: for example, *sharking* in Birmingham student jargon means 'going out on the lookout for the opposite sex', and the less regional *wicked* is a positive evaluation, as in:

> 'Yeah. She gave me roast lamb, which was wicked…'

Many other markers of student peer group membership are apparent, such as the ubiquitous use of *like*:

d it's not working. Well, I can put	like a quid to it but I didn't have enou
there and then I had a gin and I was	like a bit but I wasn't as bad as you 'c
as no-one gets offended it's just a	like,an impersonal… Ja: I just can't
it and, this sounds really bad, but	like Ann just can't go into her room and
just think it was brilliant. It was	like "Are you going to come for a drink"
im about half an hour ago and it was	like "Are you really annoyed?", and I sa
to the other one and I can talk to,	like both. J: I think you just set off o

e ripping to shreds all these really like classic songs. N: Yeah. Like turni
le and they just hit it off and just like come round and talk and it was like r
ere it's obvious that I want to just like,come in here because I should be in
becoming more and more and more like couples everywhere but I don't know,
use it at about 2 o'clock and it was like dead. They'd cut it off. So Nigel c

or the prefacing phrase, *it sounds X but,* which requests a sympathetic ear:

> 'It sounds awful but I get a feeling, it's kind of like — girls talk which…'
> 'I know it sounds really awful but if I want to stay in for the evening I can'
> 'This sounds really bad, but like Ann just can't go into her room and work'

or the framework *(it+[be]) really X 'cos,* which evaluates the coming story:

> 'It was really bad 'cos I was just like walking past…'
> 'I really liked it over the summer, 'cos like me and Andy…'
> 'the night before it was like really good 'cos when (they) are in the house…'

Student language represents a fascinating variety of spoken grammar, lexis and phraseology. We hope to make this and other data in the new corpus available to other researchers when fully corrected.

2.1.2 Other Corpora

We have also constructed many types of corpora in collaboration with outside partners. Typically, these are small and specialised.

2.1.2.1 The 'SHAPE' Corpus

One such was the one million-word SHAPE (Supreme Headquarters Allied Powers Europe) Corpus, of the English heard and read by NATO employees, both conversational and technical. This was created to support the teaching activities in SHAPE Language Centre. It is the sociolect of a close community, reflecting its concerns. The texts cover the whole range of language parameters. Compare the following extracts:

> 'It is a basic tenet of the Alliance that each member nation is responsible for the continuing support of its forces…';
>
> 'The Italian NMR requests that all Italian NCOs, ORs, and Carabinieri be excused from international duty from 2000 hours to 2400 hours on 2 October 1987 in order for them to attend a national reception';
>
> 'SHAPE CINEMA: Dress Requirements: Individuals wearing dirty,

> greasy or smelly clothing will be denied entry. Clothing designed as under garments may not be worn as outer garments…’
>
> ‘CB: Hi Barbara . this is Colonel B . How’re you doing today
> BS: Oh fine thanks . and you
> CB: Say I was wondering if you had any offers on the car you had for sale…’.

Both British and American English varieties are represented in the data, as is non native-speaking English. It has proved to be a valuable resource for SHAPE, both in the production of tailor-made EFL courses, and as an item bank for SHAPE language tests, which have been devised by testing experts from Reading University.

2.1.2.2 Corpora of examination papers

Another category of specialised corpora that we have been developing since 1986 is that of examination papers, primarily for English language. Some of this has been done in collaboration with boards that examine English as a Foreign Language. The process involves the complex task of multi-level coding of the language content of rubric within the texts, a procedure that has been discussed in a brief paper (Renouf 1988). An evaluation of examination stimulus material and rubric without recourse to the student responses that they evoke is for some purposes incomplete, and we have also analysed student scripts on occasion.

Simple concordances and wordlists can focus the user on aspects of language that might otherwise be taken for granted. Take the example of the use of the word YOU in the rubric of an examination for native speakers:

```
    t the plan will not be marked.  You  should write between 350 and 600 words.
   red Place. Write about a place   you  know which has a special atmosphere ab
    is punished by her father. (a)  You  are Ursula’s mother. Why did you never
    a) Jot down the qualifications  you  need to become an army apprentice.
(b) ture on summer holidays there.  You  have a colleague in Italy who has sent
```

It is clear from these few lines that the referent of the word varies — between the real ‘you’, as candidate and as person; the imagined ‘you’; the real ‘you’ in an imaginary situation; and so on — and that the examinee must adopt the appropriate persona. This may well not cause problems, but it is as well for examiners to be made aware of such conventionalised and largely subconscious strategies.

Just one of the features of the rubric that we have identified as a candidate for analysis is the practice of embedding the ‘trigger’, the clause that says what is to be done, within a series of sentences that give secondary instructions, advice, information about the circumstances of the text, and so on. An example is:

> SECTION B
>
> You are advised to spend approximately one hour on this section.
>
> When you have read the information given about the people in ALL FOUR EXTRACTS answer BOTH questions which carry equal marks. These questions are based on the material in SECTIONS A and B.
>
> Your answers must be based on the information given. Choose the details which you consider relevant to each of your answers and express them in your own way as appropriate.'

In this extract, 'answer BOTH questions' is the actual instruction. We refer to 'pre-instruction' text as 'preamble', and informally, by analogy, the subsequent components have been dubbed 'amble' (the clause or sentence carrying the main instruction) and 'postamble' — which terms may or may not make their way into the language eventually! The 'amble' may prove to need unearthing in some cases.

Papers of subjects other than language are also coming under scrutiny, in order to identify some of the features of language formulated by non language-specialists, and to compare the problems of readability of 'content-based' and 'language-based' examination papers.

The ultimate purpose of the study is to discover facts about this important 'genre' of language. More practically, we hope to facilitate the process of editing draft examination questions and training new question writers, and to move towards an eventual standardisation in rubric formulation.

2.2 Corpus–based Research

2.2.1 The AVIATOR Project

Our interest remains in the study of corpora of all types, but we have been conscious for some years that a static corpus is by definition fixed in time, and allows only a synchronic study, a snap-shot, of language which is in fact constantly changing.

Recently, we have been successful in securing government funding for a large, three-year project to study those changes, as reflected in a flow of data, a 'dynamic' corpus. The project is known as AVIATOR (Analysis of Verbal Interaction and Automatic Text Retrieval). It has two industrial partners, Nimbus Records and Collins Publishers, and the project team was set up last autumn.

One aim of the project is to develop types of text-searching software that will automatically monitor the changing lexical inventory of English.

2.2.1.1 New Words in English

The software designed to find new words is being developed, and new words are beginning to emerge. Of course, there are degrees of novelty, and at the moment we are erring on the side of capturing everything that is new in the newspaper text that we are looking at.

Some of the words are becoming established already, such as:

went on a coach trip to an	ACID-HOUSE	party. It was a view of Brit
have heeded the government	DRINK–DRIVE	advertisement campaign have
ave eaten food that's been	MICROWAVED	I have been sick afterwards

Some of the words are the product of perfectly respectable word-formation rules, although unlikely to become mainstream usage, such as:

te coats, nor the men like	DANDRUFFLESS	barristers. There is a sort
ifornian walnuts, and as I	DETROLLEYED	this onto the counter, some
tell. Washed with oriental	SCRUPULOSITY,	undergarmented against und

In due course, we shall know whether that particular prediction is true, since we are committed to monitoring the comings and goings of all the words that we encounter over the next period of years. It will be interesting to see which of the following new inflexions and words, for example, have staying power:

be accused of Glasgow 1990	BANDWAGONISM.	Ward refutes any such all
t that the county had been	GRANT-CAPPED	over several years and this
rs. There is no feeling of	GUNG-HO-NESS	on either the sales or prog
ch can cause anglers to be	HANDBAGGED	by swan lovers) to reels
and girls. There were hints of	HOMOSENSUAL	experiences that make one th
opening vignette on Essex	LADDISHNESS	(Walk into a pub in Hornchu
rmer prime minister of 58.	MAJORISM	may not quite have run its cou
ing to win in its last two	MEGABIDS:	for Pilkington in 1987 and, e
5 to 64. It concludes that	MIDLIFERS	are at the peak of their earn
gest worry. To placate the	PEACEMONGERS	he has agreed to talks with
itish religious elite, the	RELIGENTSIA,	can only sneer that such d
developments. England have	SEMAPHORED	their state of desperation ac
minent initially with some	THRUSTFUL	running but Harrogate led when
s that are already heavily	TOURISTED,	with mixed feelings; but this
re confident once they had	WHEEL-CLAMPED	Norwich's efforts.
claim and counterclaim of	'YUPPIFICATION'	and 'sell-out'.

The diachronic and quantitative view of language that we are taking will ultimately identify the dominant trends in word formation in

English, of both journalistic and other varieties.

2.2.1.2 New Word Combinations and Uses

We also hope to devise an automatic procedure for identifying new word combinations, meanings and uses, by studying the collocational patternings of words. We would like to notice that two established words, such as *sleeping* and *policeman* are beginning to occur together; to see that *tuna* and *yoghurt* (where *tuna yoghurt* is a new type of food product) are starting to co-occur. There is also the question of whether and how to monitor the very common words, to see, for instance, that *get* and *in* (where *get–in* is a new noun compound referring to the delivery entrance for lorries at warehouses, etc) are appearing together. As far as new meanings go, the word *floating* will need to be identified where it occurs in the context not of boats but of the stock market; and *mouse* not of *cheese* but of computers.

2.2.1.3 Text Retrieval

Another goal of the AVIATOR project is to make a contribution in the area of text retrieval, to devise a more sophisticated method of finding relevant texts in a large database than simple key-word search.

2.2.2 The Automatic Abridgement Project

This project is based on the work of and co-directed by Dr Michael Hoey (1991), and involves us in implementing some aspects of his manual system of text abridgement by computational means. Hoey's system has been demonstrated to work manually on non-narrative text-types. The methodology involves the identification of key sentences in text as those which are 'linked' and 'bonded', on the basis of patterns of lexical repetition at word and sentence level, most heavily with other sentences. When extracted from the text and juxtaposed, these key sentences together summarise the text and are internally cohesive and comprehensible as a text. Our task is to develop a system whereby the computer can automatically recognise the relevant instances of repetition, and produce the abridgement for any of the texts held in a large database.

Exact repetition is easy to identify; lemmatised repetition is less straightforward, depending on how rule-based the system is to be, and how dependent on the slower look-up procedures. We have already completed this stage. The next stages, of identifying sense relations, and of paraphrase, present a real challenge.

2.3 Collaboration on Other Projects

The Unit sometimes has an opportunity to share its corpus-building and processing experience with others. At Birmingham we have, for example, recently taken on the task of designing the recorded speech component of the new ‘Bank of English’, mentioned earlier. Outside Birmingham, the Unit has helped in a small way in processing data for established corpus-based projects such as those at the Universities of Queens, Belfast, and Stockholm (Ljung, 1990).

2.4 Corpus–Based Teaching

Finally, teaching. Corpus-based teaching is, not surprisingly, on the increase at Birmingham. CALL and data-driven learning are the specialities of EOSU (the English for Overseas Students Unit), as said earlier, but our Unit has also initiated courses in corpus linguistics.

The resources required to run a corpus linguistics course can be considerable. If, as in our case, the purpose is for students to carry out a mini-study of some aspect of language with reference to on-line concordance data, it is necessary to establish the requisite corpus access for about twelve people. Last term, we spent many days coaxing four different corpora (three written corpora of one million words, and one spoken corpus of half a million) onto a cluster of micro-computers, as our basic resource. The classes have to be double staffed, with one linguist/teacher and one computer expert present to meet all eventualities. Practical constraints mean that attention is best focussed on words of middle-range frequency; the common words generate too much data to handle easily, and the rare words make for an unrewarding search in these small corpora.

However, the reward comes in seeing teachers become researchers, and gaining a fundamentally different perspective on language. Current studies are yielding interesting facts, in grammar on such topics as ergativity; in lexis on idioms, discourse items, confusables, and the restriction on numbering in English. The corpora, now installed, will remain as a Faculty-wide resource.