Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching

A paper presented at the First international conference: Practical Applications in Language Corpora (1997)
University of Lodz, Poland
For more information on the Proceedings of this conference contact Jim Melia (zajeanfa@krysia.uni.lodz.pl

Christopher Tribble, April 18, 1997

Contents

1. Introduction

My main purpose in writing this paper is to show that it is possible to begin to use a "data-driven" approach (Johns T 1991a) to language learning and teaching even if you do not have access to established corpus resources. A secondary purpose is to discuss the potential of small, very specific corpora for ELT (Aston G, 1995) contrasting them specifically with the moderately-sized balanced corpora discussed by Biber D, S Conrad & R Reppen, 1994. My starting point, therefore, is that although I have long advocated the usefulness of corpora as a resource for the language teacher and language learner (Tribble C. 1989; Tribble C. 1991a.; Tribble C. 1991b; Tribble C. forthcoming 1997; Tribble, C & G Jones. 1990, 1997), I do not feel that these resources have to be the latest, biggest corpus. As the title of this paper suggests, I am proposing that small, informally produced corpora can be a useful resource in the language learning / teaching project. Hence the "quick-and-dirty" of the title.

Current corpus resources

If corpora and corpus-based exercises are useful because they "favour learning by discovery - the study of grammar (or vocabulary, or discourse, or style) takes on the character of research, rather than spoonfeeding or rote learning" (Tribble, C & G Jones. 1990:12), what sort of corpus is likely to be best suited to help learners of English? In the next part of this paper I shall discuss the main characteristics of some of the major corpora which are currently available, and will outline principles which have informed their development

Historically, the major trends in the development of modern electronic corpora can be traced from the Brown corpus of American English (Kuera H & W N Francis 1967) and the Lancaster, Oslo, Bergen (LOB) corpus of British texts (Johansson S 1980). These aimed to be *representative* corpora; that is to

say that the criteria used for text selection were set so as to ensure the best possible representation of a range of domains, genres and text types, and the texts themselves were restricted to a maximum of 2000 words per item in order to maintain this balance. Some major corpus projects such as the British National Corpus (BNC) (Burnard L 1995) - a 100 million word representative corpus of contemporary British written and spoken texts stand in direct line of succession to Brown and LOB. The development of "monitor corpora" such as the Bank of English at Birmingham University (Sinclair J 1991) represent a different approach. Here it is assumed that the corpus "has no final extent because, like the language itself, it keeps on developing" (Sinclair J 1991:25). Standing apart from this mainstream, though often drawing on the design principles of the Brown and LOB tradition, other types of balanced corpus are also emerging - the Louvain Corpus of Learners' English (Granger S & S Tyson 1996) is a good example of a corpus developed for interlanguage research.

Whatever the strengths or weaknesses of the approaches adopted by the compilers of these earlier corpora, and however great the contribution they have made to the development of new grammars and dictionaries of English, the corpora themselves have not been easily available to or widely used by most language teachers. The reasons for this are manifold - commercial, technical, legal - but a major issue has always been that even if teachers had wanted to use these corpora, and if the had had the right so to do, such text databases have, so far, been too expensive or too technically demanding for the computing resources that most foreign language teachers command.

It might seem that this problem is disappearing as the rapid development of telecommunications for computing means that now (or very soon) a large number of teachers and students will be able to access the BNC or the Bank of English on-line and use the same search engines as their university or commercial counterparts. While this may be true, I would hold that this will not meet all the needs that language teachers have for corpus resources. Firstly, with regard to the question of access, I would accept that the on-line availability of major English language corpora could have a significant impact on a growing number of teachers in some countries. I would also say that for many others access will remain a problem and that for a number of years to come not everyone will have the budget, the computers, or the bandwidth to work with a corpus this way (especially in state sector secondary schools). Secondly, such large corpora can present significant methodological problems for less experienced users - the risk of drowning in data is high! Thirdly, even if students and teachers could have unlimited access to multi-million word corpora, I am not convinced that these will necessarily be the resource best suited to meet their needs. My concern here relates to a problem of authenticity.

Widdowson comments on the notion of "authenticity" as follows:

> "An authentic stimulus in the form of attested instances of language does not guarantee an authentic response in the form of appropriate language activity .... we should retain the term 'authenticity' to refer to activity (i.e. process) and use the term 'genuine' to refer to attested instances of language (i.e. product)" (Widdowson HG 1983:30)

I feel that this distinction can also apply in the context of corpus applications in language teaching in the sense that, although a corpus may contain millions of "attested instances of language", there is nothing to guarantee that you can use data from that corpus as a stimulus for "appropriate language activity". EFL students are unlikely to be motivated by a language learning activity if the instances of language use that they are studying are taken from contexts which make no connection with their interests and concerns. It may well be possible to use either a corpus of telephone engineering texts or the BNC itself to provide an elegant account of the rules for definite article use in written text. It may be extremely difficult, however, to persuade students to engage with this account if the contextualising instances do not engage them in some way other than the purely analytic. *Genuine* examples of language in use will not necessarily lead to *authentic* language use or effective language learning activities.

Which corpus?

What sort of corpus do foreign language learners and teachers need then? I would say first of all that they almost certainly need many corpora rather than one. Students of English want to be able to

write different kinds of text and become effective language users in many different contexts. They need a rich set of potential models for their own language behaviour. As Flowerdew says:

> "Many native speakers make use of others' writing or speech to model their own work in their native language where the genre is unfamiliar. It is time that this skill was brought out of the closet, and exploited as an aid for learning." (Flowerdew J 1993:309).

A *range* of corpus resources can provide the different resources that learners need . But again, what sort of corpora? Where can learners get hold of the authoritative models that they need?

The notion of the authority of a corpus is associated with the issue of authenticity mentioned above. For the developers of Brown, LOB and the BNC - driven mainly by an interest in grammar, lexicography and natural language processing - an authoritative corpus had to be both extensive and balanced in terms of content, genre and text length. The builders of monitor corpora (such as the Bank of English) appear to feel that balance has become less of a priority - sheer size seems to have become the basis for the corpus's authority. Being concerned with the problem of learner writing, those involved in the Louvain Corpus of Learners' English, argue that a corpus of native speaker texts which are directly analogous to the essays they hold in their learners' corpus is the most authoritative collection for their purposes. My own view is that the most useful corpus for learners of English is the one which offers a collection of *expert performances* (Bazerman 1994:131) in genres which have relevance to the needs and interests of the learners. Collections of relevant expert performances will exemplify the results of the desired forms of language behaviour that learners are trying to achieve and will also constitute motivating starting points for language learning and language using activities. Thus it may be that a mixed corpus of literary and journalistic texts will be attractive to general English learners. It may also be the case that another group of students will be delighted to find themselves working on a corpus of telephone engineering texts. It will depend on what they are interested in. Those students who are beginning to learn how to write in formal, professionally oriented contexts need a different sort of corpus.

The non-standard corpus which I am advocating for students with an interest in learning how to write factual texts can be drawn from a source of texts which exists in many language schools or English language departments, but which is not usually used as a *language* learning resource: it is the widely available multimedia encyclopedia. In this paper I shall give examples from Microsoft Encarta® 96 - World English Edition (Microsoft 1996). This is not an endorsement of this particular product: I have chosen to use it as an example of what can be done with this sort of text source because it is widely available, and because I have a copy to hand. Other multimedia encyclopedias (such as the Hutchinson® or Grolier®) offer equivalent functionality and quality of data.

Several arguments can be raised against a corpus built with this sort of data:

1. It will always be restricted in size compared with the major corpora such as BNC
2. It will be unrepresentative in the range of content that it covers
3. It will contain an unrepresentative range of written genres,
4. Its authorship will be unusual and unrepresentative
5. It will contain texts which are in themselves an unusual genre
6. It is an inappropriate set material for EFL students

Remembering that I have already stated that I am proposing the use of such a corpus with *students who are beginning to learn how to write formal, professionally oriented texts* I feel it is possible to counter each of these objections.

With regard to the size of the corpus, although Microsoft gives no specific indication of the number of words in the encyclopedia, there are 27,033 separate articles listed in the "search" facility. These contain anywhere between 200 and 5000 words. If we take a pessimistic mean of 1000 words, then the biggest possible "corpus" based on this source of texts could contain 27,033,000 words. A not insignificant number - to quote a much parodied recent British Prime Minister.

Likewise, when it comes to balance of topics, these 27,033 articles are divided into reasonably evenly distributed domains:

Physical science	3245
Life science	3509
Geography	6047
History	4706
Social Science	2911
Religion and Philosophy	2150
Art, Language and Literature	3588
Performing arts	1823
Sports, Hobbies and Pets	1006

While these topics do not cover areas such as current events, or contain examples of strongly opinionated argumentative writing, they do map quite closely on to the main subject divisions which are found in most secondary education systems. When the breadth of sub-topics in particular domains is considered, then the range is impressive. A corpus based on this data should certainly contain enough texts which *most* students in *most* language classes will find interesting and informative.

The question of genres and text types and overall level is more difficult. However, if again one considers the sorts of texts which are significant in national modern language examinations or internationally recognised language tests, the encyclopedia does contain a significant set of core text types and genres. These include:

- Essay descriptive / discursive
- Process descriptions
- Physical descriptions
- Biographies

While, therefore, a corpus based on a CD-ROM encyclopedia will not include examples of contemporary correspondence or fictional writing, it will contain a sufficient range of text types to be useful, and - as a quick-and-dirty solution - will give teachers and students a more coherently structured and usable resource than they will get from any other cheaply and easily available resource currently available.

Two final issues remain - the first is the "unusualness" of encyclopedia entries; the second their relevance for learners of English. The first problem is best addressed by looking at the texts themselves. The example below (the opening of a 260 word essay) shows the potential drawback of using encyclopedia entries as a corpus resource. The problem lies in the opening sentence where the editorial style of the "super-genre" ENCYCLOPEDIA obliges the author to front the name and dates of the subject of the essay and to drop the article and copula verb in the opening sentence. The good news is is that the text can be "normalised" by the addition of "was the" immediately after Tambo's dates. This done, the text reads naturally and represents a realistic target for an upper intermediate to advanced learner of English, in terms both of style and extent.

Tambo, Oliver (1917-1993), South African political leader, who, while in exile from 1960 to 1990, led the African National Congress (ANC), then an illegal organization. Tambo, along with other black leaders of his time, opposed apartheid, the South African government's strict policy of racial segregation. Born in Bizana, Transkei (now South Africa), Tambo received a scholarship to the University of Fort Hare, from which he earned a B.Sc. degree in 1941. Tambo stayed on at Fort Hare to gain a degree in education, but was expelled in 1942 for leading a student strike.

The last problem - the relevance of the texts to the interests of learners of English and their overall level - is also addressed by the Tambo example which is, I would contend, sufficiently informative to be worth reading, and which does not impose excessive demands on a reader who has an *interest* in reading it. After all, it was initially written to be used by school students in English speaking countries.

2. Using a CD-ROM encyclopedia micro-corpus

If, then, we accept that this sort of CD-ROM encyclopedia might be a useful source of corpus data for a teacher of writing with students at upper-intermediate or higher levels two immediate questions arise. "How do you construct the corpus?" and "what can you do with such a corpus once you've got it?"

Assuming that you have access to an installed copy of Encarta® and that there is some sort of wordprocessor on your PC, the procedure for building a micro-corpus is straightforward:

1. Open your word processor / open Encarta®
2. Use the Encarta Pinpointer to identify texts associated with a topic area in which your students are interested or have a learning need. In this example we will assume that the theme is "business"
3. When you have found a text you think will be useful, select "*copy*" from the menu bar at the top of the article
4. Copy the article to the Windows clipboard
5. Switch to your word processor (Ctrl+Tab in Windows 3.xx) and paste the article into an empty document
6. Edit the first line to remove "unusual language"
7. Save the text as a "plain text" or ASCII / ANSI file into an appropriate subdirectory and with a useful file name (e.g. profit01.bus)
8. Switch back to Encarta® (Ctrl+Tab in Windows 3.xx) and search for more texts.

Working in this way, I have been able to construct themed, twenty to thirty-thousand word micro-corpora in fifteen to twenty minutes. Although such a corpus sounds insignificantly tiny when compared with the huge corpora which can now be accessed, I would argue that if one wishes to investigate the lexis of a particular content domain (e.g. health) a specialist micro-corpus can be more often be useful than a much larger general corpus. For example. in the written component of the BNC Sampler (1,000,000 words) there are no instances of "cancers". An Encarta® micro-corpus of health articles (24,805 words) gives 33 usefully contextualised examples!

The way you will use these micro-corpora will depend on the interaction between teaching and learning purposes. In his opening plenary at this conference Michael Hoey discussed the range of questions one can ask about a word. These included:

1. What lexical patterns is the word part of (collocation)?
2. Does the word regularly associate with other meanings (the notion of *semantic prosody*)?
3. What structures does it typically appear in (colligation)?
4. Are there any correlations between the word's uses / meanings and the structures in which it participates (colligation)?
5. Is the word associated with any position in text organisation?

These same questions can be asked of words or phrases in an encyclopedia micro-corpus - and asked in contexts which hold some interest or relevance for the student. A summary of some of the different uses to which micro-corpora can be put is presented below, and give an impression of the potential value of such resources. This framework for corpus investigation assumes the availability of one of the new generation concordancing / wordlisting programs such as WordSmith Tools (Scott M 1997) or MonoConc for Windows (Barlow M 1997), and a word processor with search / word count facilities for the investigation of whole or part texts and the preparation of teaching materials.

3. Using the Texts - a provisional framework

Single theme micro-corpus

Area	Specific to?	ELT Focus / Value
High frequency lexis in a specific content domain	Topic area and Text types	• vocabulary enhancement in familiar content area • consolidation (understanding / use) with texts that are relevant to interest / current learning focus

Collocation / Colligation Semantic prosody	Topic area and Text types	• enhance SS awareness of collocation / colligation of *known* words in context expose SS to *new* words associated with *known* words in familiar content area
Grammar and discourse	Text types and Genre	• investigation of grammar issues in familiarlexical domain (full text / part text) • investigation of discourse organisation in appropriate text types (full text / part text)

Multiple theme micro-corpus

Area	**Specific to?**	**ELT Focus / Value**
Grammar and discourse	Text types / Genre	• investigation of grammar issues in familiar lexical domains / genres • investigation of discourse organisation in appropriate text types
Contrasts in lexical use	Topic area and Text types	investigating the meaning / behaviour of words in different domains.

Single text type micro-corpus

Area	**Specific to?**	**ELT Focus / Value**
Grammar and discourse	Text types / Genre	• investigation of lexical patterning in single text types / genres • investigation of discourse organisation in single text types / genres (whole text)

Although space does not allow a full demonstration of the sorts of materials which can be developed from such resources, a small example still indicates the potential value of these micro-corpora.

4. An example

High frequency lexis in a specific content domain - with thanks to Tim Johns for the idea of looking at "such as" (see Tim John's Home Page at http://sun1.bham.ac.uk/johnstf/homepage)

Teaching context	Upper intermediate students with an interest in business communication.
Resources	An Encarta micro-corpus / A single PC + printer with MonoConc for Windows or WordSmith Tools or similar concordancer / a means for reproducing teaching materials
Teaching purpose	To investigate superordinate / co-hyponym relations in Business and Economics writing.
Procedure	Use concordancer to identify instances of "such as" delete non-useful examples save as text file (allow up to 600 characters of context) use a word processor macro to select complete sentence contexts and edit the text to create appropriate blanks and prompts

Sample data:

```
es to meet its needs. Other activities, such as manufacturing, were regarded as
 proliferation of free-trade agreements such as those under the General Agreemen
" and added Communist-type appellations such as "People's" or "Popular" to their
tries are supervised by a central bank, such as the Bank of England in the Unite
pon Credit Bank Ltd. Some private banks such as Dai-Ichi Kangyo Bank (the world'
re they are owned by the nation's banks (such as the United States and Italy), t
hat cannot be easily converted to cash, such as buildings and equipment, are con
hat can be converted readily into cash, such as finished goods or stocks and bon
ted and paid more attention to concepts such as quality, value, and usefulness.
may be items for immediate consumption, such as a TV dinner or a bar of chocolat
```

```
rocedures. Nevertheless, many countries (such as the United States and Great Bri
 owners' equity as of a particular date (such as the last day of the accounting
as well as details about long-term debt (such as interest rates and maturity dat
 radio, and television, special devices such as single-product folders or multip
```

Teaching materials

Choose the appropriate superordinate for the gap in each of the sentences given below.

1. Noncurrent ---------- are usually debts that will come due beyond one year-*such as* **bonds, mortgages**, and **long-term loans**.
2. The direct marketing of ---------- *such as* **cosmetics** and **household needs** is very important.
3. Often the manufacturer must provide ---------- *such as* **installation** and **maintenance** for a specified time period.
4. On the consumer level, sales promotion may involve special merchandising ---------- *such as* **discount coupons**, **contests**, a **premium** with the purchase of a product, or a **lower price** on the purchase of a second item.
5. Similarly, the purchase of durable or long-lived ----------, *such as* **refrigerators**, **automobiles**, and **houses**, may be deferred when the economy is declining and may increase rapidly in periods of prosperity.
6. Staple ----------, *such as* **food** and **clothing**, tend not to be seriously affected by the business cycle.

goods / inducements / liabilities / products / products / services

5. Conclusion

In presenting this paper I hope that I have done more than provide a recipe for using a particular reference resource for language teaching. Certainly, I shall be pleased if those who were sceptical about the potential value of using a CD-ROM encyclopedia as a source of corpus data have been persuaded of the value of this sort of text (and I think that the natural quality of the examples in the exercise above will persuade most readers of this). I shall be even more satisfied, however, if what I have said here helps teachers to start using the vast range of language data which now exists in electronic form *now* rather than feeling that they must wait until they have access to a "real" corpus.

At this year's (1997) IATEFL Conference in Brighton, UK, it was clear from many presentations - including two conference plenaries - that the corpus is no longer the sole preserve of the university or commercial research team. Teachers and students are beginning to have access to corpus resources and are beginning to work with them in interesting and creative ways. Most of the corpora which were discussed at IATEFL had been developed for specific purposes - usually lexicography. I hope that this paper - and others which have been presented at PALC -will serve to remind teachers and students of the many other sources of language data which exist outside these established corpora. I also hope that it will stimulate debate around the issue of *which* corpora learners need - what are the right models for specific learners with specific (or general) needs? Are L1 essays the best model for L2 apprentice essay writers, or would (along with the sort of encylopedia texts discussed in this paper) the expert performances found in newspaper editorials be a fruitful source of models? Do writers of scientific texts always need to work from other scientific texts as they develop their capacity as writers. Probably yes, if we consider their need to develop an understanding of the structure of specific genres. Perhaps no, if we think of the scientist's need to develop a metaphoric language - it is possible that novels and newspapers will give them more of what they need in this instance. I am sure that the debate will continue!

Bibliography

Aston G (1995) "Corpora in language pedagogy: matching theory and practice" in Cook G & B Seidlhofer (eds) Principle and practice in applied linguistics: studies in honour of HG Widdowson OUP

Barlow M (1997) *MonoConc for Windows* Athelstan, Houston TX

Bazerman C (1994) *Constructing Experience* Southern Illinois University Press Carbondale

Biber D, S Conrad & R Reppen (1994) "Corpus-based approaches to issues in Applied Linguistics" Applied Linguistics 15/2:169-188 OUP Oxford

Burnard L (1995) T*he British National Corpus Users Reference Guide* (SGML version) Oxford University Computing Services Oxford

Encarta '96 (1996) World English Edition Microsoft Home

Granger S & S Tyson (1996) "Connector usage in the English essay writing of native and non-native EFL speakers of English World Englishes 15/1: 17-27

Flowerdew J (1993) "An educational, or process, approach to the teaching of professional genres" ELTJ 47/4 305-316: 305-316 Oxford University Press Oxford

Hoey M (1997) *From concordance to text structure: new uses for computer corpora* paper presented at the 1997 PALC Conference, odz, Poland

Johansson S (1980) "The LOB Corpus of British English Texts: Presentation and comments" ALLC Journal

Johns T (1991a). "Should you be persuaded---Two Examples of Data-Driven Learning Materials" *English Language Research Journal* (4) 1--16. University of Birmingham

Kuera H & W Nelson Francis (1967) *Computational analysis of present day American English* Brown University Press Providence Rhode Island

Scott M (1997) *WordSmith Tools* OUP Oxford

Sinclair J (1991a) *Corpus, Concordance, Collocation* OUP Oxford

Tribble C. (1988) The use of text structuring vocabulary in native and non-native speaker writing, MUESLI News, June 1989.

Tribble C. (1991a) Some uses of electronic text in English for academic purposes, in J. Milton & K. Tong (eds.) Text Analysis in Computer Assisted Language Learning. The Hong Kong University of Science and Technology: Hong Kong, 4-14.

Tribble C. (1991b) 'Concordancing and an EAP writing program', CAELL Journal 1/2, pp. 10-15.

Tribble C. forthcoming (1997) 'Corpora, Concordances and ELT' In Boswood T (ed) *New ways of using computers in language teaching*, TESOL, Alexandria VA

Tribble C & G Jones. (1990. *Concordances in the Classroom: A Resource Book for Teachers*. London: Longman.

Tribble C & G Jones. (1997) *Concordances in the Classroom: using corpora in language education* (new edition) Athelstan, Houston TX

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