The Corpus Revolution in EFL Dictionaries

Ramesh Krishnamurthy

1. Introduction

The early history of monolingual EFL dictionaries was described in detail by A.P. Cowie in Kernerman Dictionary News (2000). In his article, Cowie said: "And the authenticity of the grammatical claims made about English, and of the examples selected, has been improved beyond recognition by the use, since the early 1980s, of large-scale computer-stored corpora of English, the best known of which are the British National Corpus and the Bank of English."

This paper will describe the revolutionary impact that the use of large computer-held corpora has had on EFL dictionaries since the 1980s, with special reference to the Cobuild project (home of the Bank of English corpus) at Birmingham University, which in many ways pioneered the developments. As Michael Lewis has said (2001): "The first Cobuild dictionary changed the face of dictionary making, and the way some of us thought about vocabulary, for ever."

2. Traditional sources of lexicographic evidence

For centuries, lexicographers had to rely on their own and their colleagues' intuitions and language experience as the basis for their descriptions of language. They also frequently made use of descriptions in previously published works, thus perpetuating any errors and inaccuracies.

However, individual intuition and experience are subject to limitations. As John Sinclair has said: "Users of a language are not necessarily accurate reporters of usage, even their own" (1987); "Using a language is a skill that most people are not conscious of; they cannot examine it in detail, but simply use it to communicate" (1995); and "There are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently" (1995). Even highly-skilled, highly-trained, and extremely dedicated lexicographers inevitably attain only a partial knowledge of a language. They also suffer from the general human weakness of a poor or selective memory. And, of course, lexicographers' work is affected by their own prejudices and preferences, however subconsciously.

One way to lend more authority to intuition-based dictionary entries is by adding authentic citations as evidence. Two historical English dictionaries are particularly noted for adopting this policy. Dr. Johnson's Dictionary (1755) deliberately took its citations only from "the best authors" writing in "the golden age of our language", and the citations therefore reflect only the higher culture. Furthermore, Johnson frequently altered the original texts to suit his purposes, for example quoting the same line from Milton's Paradise Lost with "outrageous" at one entry and "outragious" at another, and the same line from the Bible with "**in**discreet" and "**un**discreet", etc. (Kwon 1997). The Oxford English Dictionary (OED, 1879-1928) covered a wider range of authors and texts, but still managed only a piecemeal coverage, because the editors discovered that readers asked to select examples from texts tended to notice the unusual items and overlook the commonplace[1].

3. The corpus as lexicographic resource

John Sinclair has compared the impact of corpora on linguistics with that of telescopes on astronomy. The use of corpora is rapidly changing our ideas about language, and corpus research has already revealed that many of our past intuitions were wrong.

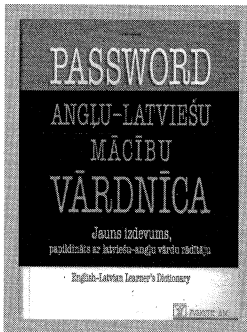
A large computer-held language corpus can overcome many of the limitations of human linguistic intuitions. It can be far more comprehensive and balanced than any individual's language experience. It does not have any memory problems, and can immediately recall all the information that has been input. It does not get distracted by unusual items, but can show us both what is common and typical and what is rare or restricted in use. Ultimately, the corpus can provide more objective evidence.

Further inadequacies of human linguistic informants have come to light: we cannot quantify our knowledge of language[2], we cannot invent natural examples[3], and we are unable (especially since the advent of the Internet) to keep up with language change. Corpora are able to assist us in all these areas: they can give us accurate statistics, a vast number of authentic examples, and (if frequently updated) can reflect even very recent changes in the language.

Another objection to using the intuitions and experiences of one individual is that they can easily be challenged or refuted by others. Corpus data encompasses the language use of many members of the language community, and therefore carries

Ramesh Krishnamurthy has degrees in French and German from Cambridge University, and in Sanskrit and Indian Religions from the School of Oriental and African Studies, London University. He has worked on the COBUILD project at Birmingham University since 1984, contributing to several dictionaries, grammars, and other publications, as well as developing corpora and software. He is an Honorary Research Fellow at Birmingham University and Wolverhampton University, has taught and supervised MA and PhD students, participated in EU linguistic projects, and conducted workshops and courses in several countries.
www.ccl.bham.ac.uk/ramesh

An appendix containing raw corpus data and the author's analysis of *sexy* is available online:
http://kdictioneries.com/newsletter/kdn10-sexy.html



PASSWORD
English-Latvian Learner's Dictionary, Second Edition
Zvaigzne ABC
Riga, 2002
Translation:
Maija Treilona and
Rasma Mozere
Editor: Ieva Tarvida
ISBN 9984-22-075-3
992 pp.
25 x 170 x 55 mm.
Hardcover
http://kdictionaries.com/products/medium/medium-latvian.html

greater authority. Language corpora also represent the democratization of the sources of evidence. We may be able to criticize Johnson's limited range of carefully-vetted sources, and even the wider but only partly used range of OED texts, but it is difficult to argue with evidence of language usage that is repeated by hundreds or even thousands of different speakers and writers in a variety of situations and contexts. In addition to the literary canon, corpora include tabloid newspapers, popular magazines, and recordings of informal conversations.

Finally, every language has its cultural connotations and underlying ideologies, which are difficult for individuals to perceive. The corpus can be invaluable in revealing these[4].

There were some problems with the use of corpora until the 1980s. Very few corpora were available, and they were too small for most purposes (the largest was around 1 million words). They were able to provide only superficial indications about many linguistic features, and were reliable only for the most frequent words in the language (i.e. grammatical words). As larger corpora were built from the 1980s onwards, attention turned to the question of balance: what proportions of texts from which genres should be included? The earlier problems of the non-availability of data, and the technical difficulties of converting printed and spoken texts into digital files had been resolved. But we were now faced with the sudden superabundance of digitalized journalistic texts, especially newspapers.

4. Earlier EFL Dictionaries

The earlier EFL dictionaries for advanced learners (i.e. the 3 editions of Oxford Advanced Learner's Dictionary (OALD), which was the sole example of this genre from 1948 to 1974), developed mainly by language teachers, had a fairly prescriptive attitude to their audience. At that time, most students studied languages at a university, and focussed on literary, historical, and higher-cultural texts. Inclusion policy in EFL dictionaries therefore favoured literary and higher-register items over more colloquial ones.

These dictionaries were also more influenced by the native-speaker lexicographic traditions (e.g. OALD claimed that it combined "the traditions of the Oxford Dictionaries" with the "language-teaching skills" of its editor, A.S. Hornby [Preface, 3/e, 1974]). The ordering of senses initially followed native-speaker practice in putting historical and etymological meanings first. The definition style was simpler but still ters rather like the language of telegram and often included abbreviations. Son definitions closely resembled the one-wo or short-phrase synonymic equivalen given in bilingual dictionaries.

The main deviations from native-speak lexicography were the omission of mo etymological information, the marking syllable-division (or hyphenation) poin in headwords, the use of IPA symbols f pronunciation (rather than "respelling"), t inclusion of pictorial illustrations, ar the increase in grammar informatic (mainly concerning noun countability ar details of verb complementation, going f beyond the simple transitive/intransitiv labels in native-speaker dictionaries). EF dictionaries had more examples tha native-speaker dictionaries, but eschewe the use of authentic citations in favour invented pedagogic "model" examples illustrate their definitions.

OALD (1948, 1963, 1974) was belatedl joined by similar dictionaries from oth publishers: Collins (1974), Longma (LDOCE, 1978) and Chambers (1980 Longman introduced some interestin innovations: a controlled defining vocab ulary; examples based on authentic da from London University's Survey English Usage; usage notes to disambiguat near-synonyms; making many embedde items into headwords and thus easie for learners to find; and using academi terminology (e.g. "phrasal verbs" instea of OALD's "verb with a particle preposition", and "collocations" instead "words that the headword usually combine with").

5. Cobuild

The Cobuild project was set up jointly b Collins (now HarperCollins) publishers an the University of Birmingham in 1980, an led by John Sinclair, who had created an analysed the world's first spoken corpu in the 1960s. The project's declared aim were to collect and analyse a large corpu of modern English, and to publish th findings in reference books for learner and teachers of English.

Initial lexicographic analyses wer performed manually on a corpus of million words, using paper printouts frequency lists and concordances, and th analyses were first entered onto pape slips, then keyed into a computer database But computational methods were rapidl introduced into all aspects of Cobuil work. Computer-typesetting was alread established, and Longman had used computer to ensure that words in LDOCE' definitions were part of its controlle

vocabulary.

Cobuild increased its corpus to 20 million words and wrote software to allow online inspection and analysis; results were entered by lexicographers directly into the database; the computer performed various editorial checks, especially to maintain consistency and validate cross-references; progress was automatically monitored; and duplication of effort was reduced, by lexicographers being provided with completed analyses of similar words. Finally, the database entries were extracted automatically into draft dictionary files, edited online, and became input files for typesetting the dictionary. This dictionary, published in 1987, was the first to make use of computers throughout its creation.

The corpus has continued to grow since then: renamed the Bank of English in 1991, it now stands at 450 million words. The corpus retrieval software has also been substantially improved, with more sophisticated search tools, wordclass tagging and syntactic parsing, automatic analyses of collocation, and so on.

6. The impact of corpora on EFL dictionaries

The effects of language corpora were first felt in EFL dictionaries, because the smaller corpora available initially were just about sufficient for the reduced coverage of an EFL dictionary (c. 50,000 entries), but completely inadequate for a large native-speaker dictionary (c. 200,000 entries). The first corpora were exclusively monolingual and mainly English, so bilingual dictionaries benefited only marginally and had to wait until corpora were developed in other languages. The main impact of corpora on bilingual dictionaries required the development of parallel multilingual corpora, which have only started to become available very recently.

A major change had also taken place in the language learning market in the preceding decade or two: students of English were increasingly studying at language schools rather than universities, less interested in literature and high culture, and demanding instead communicative language for much more mundane and practical purposes, such as tourism and commerce. As students are increasingly exposed to unrestricted, unedited and unmediated output via the media and especially the Internet, their dictionaries need to cover much more of the lexicon, at least for decoding purposes. The corpus-based generation of dictionaries therefore became more descriptive[5].

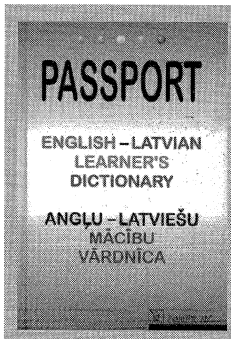
The first impact of corpora can be seen in dictionary inclusion policy. EFL dictionaries began to base their headword lists on corpus frequency, and therefore included many more journalistic and colloquial expressions (e.g. OALD6's new words: *cardboard city, generation X, latchkey child, multiskilling, outsource, innit*), leaving less space to accommodate literary and higher-register items[6]. Later editions (e.g. COBUILD2 1995, LDOCE3 1995) even published the frequency information in the dictionary itself.

Ordering of senses within entries also changed substantially, reflecting the relative frequency of the senses in corpus data, and especially the importance of hitherto largely overlooked areas of meaning. For example, COBUILD was the first dictionary to give the "homosexual" sense of *gay* first (and the "lively and cheerful" sense was labelled as old-fashioned), and the first to bring to our attention the frequent use of the verb *see* in discoursal expressions such as *I see*, and *You see*, meaning "understand", rather than in semantic meanings relating to the faculty of vision. Common verbs like *have*, *take* and *make* were seen to function in a semantically depleted way, as "delexical" verbs which merely provided the syntactic link with the following noun objects which carried the major semantic component, in phrases like *have a bath*, *take a nap* and *make a decision.*

EFL dictionaries were now able to give much better information on collocation[7], because of improved corpus software. Grammar coding became simpler, but more extensive. Wordclasses were subdivided into more subclasses, and detailed grammar patterns were given for all wordclasses, not just for verbs. And, of course, more authentic examples were supplied from the corpus data.

COBUILD in particular introduced several other major innovations: all the main forms of a headword were given in full (not abbreviated); definitions were expressed in full sentences showing typical linguistic patterns and contexts (cf "When a horse gallops, it runs very fast" with the traditional "(of a horse) to run very fast"); examples were taken straight from the corpus, with minimal editing; and, grammar and semantic relations were printed in a separate column to the right of the main text. However, unlike most of the other dictionaries, COBUILD did not use syllable markers or pictorial illustrations.

Although all of the current EFL dictionaries make some claim to the use of corpora in their compilation, they vary considerably in the extent to which they take the corpus evidence seriously. There is still heated lexicographic debate on



PASSPORT
English-Latvian Learner's Dictionary
Zvaigzne ABC
Riga, 2000
Translation:
Maija Treilona and
Rasma Mozere
Editor: Ieva Tarvida
ISBN 9984-17-232-5
650 pp.
22 x 150 x 40 mm.
Hardcover
http://kdictionaries.com/products/junior/junior-latvian.html

various issues: What is the ideal corpus? To what extent should the corpus evidence affect lexicographers' decisions[8]? Should encyclopaedic items, abundant in corpora, be included in EFL dictionaries[9]? Which aspects of the descriptive apparatus are pedagogically relevant to the student? Should authentic corpus examples be edited for pedagogic purposes?

7. Subsequent developments and the future

The influence of the innovations in EFL dictionaries is also evident in corpus-based native-speaker and bilingual dictionaries. Collins started using Cobuild's Bank of English corpus and also built its own Language Databanks for other languages (French, German, Spanish, etc). Oxford used the British National Corpus and a corpus of French in the Oxford-Hachette French Dictionary (1994), and produced the corpus-based New Oxford Dictionary of English (1998).

EFL dictionaries have seen a shorter time gap between editions (cf OALD 1948, 1963, 1974, 1989, 1995, 2000; LDOCE 1978, 1987, 1995, 2001; COBUILD 1987, 1995, 2001). This is partly due to market considerations, and partly due to the greater ease of producing new editions of computer-held texts.

Developments in computer technology have also led to the release of EFL dictionaries on CD-Rom, increasingly in simultaneous publication with the paper edition. One important benefit of CD-Roms is that pronunciations can now be heard, and do not have to be interpreted from phonetic symbols. Many dictionaries are now online; indeed, the OED has ceased paper publication and updates will only be available online from now on.

Corpus data has also become publicly available. Cobuild first released corpus data in printed form in its Concordance Samplers series, then on CD-Rom (a 5-million-word Word Bank forms part of 'Cobuild on CD-Rom'; the Collocations CD-Rom contains 2.6 million corpus examples). The availability of corpora online (Bank of English, British National Corpus, and many others, in many languages) has allowed teachers and students permanent access to native-speaker data (whereas native-speaker informants may not always be available for consultation). Researchers in CALL (computer-assisted-language-learning) are creating new software to make more use of corpus data.

As multilingual corpora become more available, and increase in size, genre variety and degree of automation, the impact on bilingual dictionaries will be immense. As monolingual corpora increase in size they will underpin most native-speaker dictionaries. Enhanced annotations of text corpora will facilitate the study of semantics and pragmatics. Improved software will deepen our understanding of collocation. Frequent corpus updates will improve our ability to identify important trends in language change. Audio and video corpora are being developed and will contribute to better information on pronunciation and intonation, on body language and gesture.

Dictionary design is one area that has been somewhat resistant to change. Even corpus-based dictionaries on CD-Rom are still rather dependent on the paper product design. Considering the exciting developments in interactive computer games, it must be feasible to create more user-stimulating language reference resources.

PASSPORT
English-Chinese Learner's Dictionary
Bookman Books
Taipei, 2001
ISBN: 957-586-945-1
685 pp.
216 x 157 x 37 mm.
Softcover
http://kdictionaries.com/products/junior/junior-chin.html

Notes

1. James Murray, the first editor of the Oxford English Dictionary, made the following complaint (1879): "The editor or his assistants have to search for precious hours for examples of common words which readers passed by... Thus, of *abusion*, we found in the slips about 50 instances: of *abuse* not five...There was not a single quotation for *imaginable*...".
2. Stubbs (1995): "Native speakers can often give a few examples of the collocates of a word ... But they certainly cannot document collocations with any thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations."
3. 'Naturalness' is a concept put forward by John Sinclair (1984), which goes beyond the earlier purely formal concepts of 'grammaticality' and 'well-formedness'. An instance of what I would consider to be a non-natural example is "Never hold a gun by the business end" (OALD 6, 2000). Apart from its pragmatic oddity (it is difficult – though not impossible – to imagine who would actually say this, and in what situation), there are no examples of *"**by** the business end"* in the current 450-million-word Bank of English corpus.
4. See, for example, Krishnamurthy (1996).
5. Even the corpus-based dictionaries could not be purely descriptive, however because of their student target audience so they still attached warning labels to non-recommended usages (e.g. in COBUILD, for the sentence-adverbial use of *hopefully*: "Some careful speakers of English think that this use of *hopefully*

is not correct, but it is very frequently used.").

6. I remember an academic review of COBUILD (1987) decrying the omission of "mizzenmast", because it occurred frequently in Herman Melville's Moby Dick, and EFL students would therefore need it.

7. However, collocations are still inaccurately reflected, even in the latest editions: for the headword *overshoot* LDOCE omits both *target* and *runway* (the most significant collocates in the corpus) and instead gives *turning* (1 example in 450m words); OALD gives *runway*, but not *target*; only COBUILD gives both *target* and *runway*.

8. Of course, dictionaries that already had pre-corpus editions faced the problem of 'text inertia': a lot of money had been invested in creating a satisfactory text from intuition, so they were unwilling to make wholesale editorial changes just because of corpus evidence; pedagogical conservatism in the teachers and students also contributed to this inertia.

9. Longman (LDELC 1992) and Oxford (OALED 1992) produced separate editions with copious encyclopaedic entries in an attempt to resolve this issue.

References

Chambers Universal Learners' Dictionary. 1980. Edinburgh: W&R Chambers.

Collins English Learner's Dictionary. 1974. London: Collins.

COBUILD: *Collins Cobuild English Dictionary for Advanced Learners*, Third edition. 2001. Glasgow: HarperCollins.

Cowie, A. P. 2000. 'The EFL Dictionary Pioneers and their Legacies.' In *Kernerman Dictionary News,* 8. (http://kdictionaries.com/newsletter/kdn8-1.html)

Johnson, S. 1755. *A Dictionary of the English Language.* London: Longman.

Krishnamurthy, R. 1996. 'Ethnic, Racial and Tribal: The Language of Racism?' In *Texts and Practices.* C. Caldas-Coulthard and M. Coulthard, (eds.). London: Routledge.

Kwon, H. K. 1997. *English Negative Prefixation: Past, Present, and Future.* Unpublished PhD thesis, University of Birmingham.

LDELC: *Longman Dictionary of English Language and Culture.* 1992. Harlow: Longman.

LDOCE: *Longman Dictionary of Contemporary English,* Third edition. 1995. Harlow: Pearson.

Lewis, M. 2001. 'Is anyone in EFL actually awake and thinking?' In *ELGazette*, 261.

Murray, J. 1879. *Address to the Philological Society.* Oxford: Clarendon Press.

New Oxford Dictionary of English. 1998. Oxford: Oxford University Press.

OALD: *Oxford Advanced Learner's Dictionary of Current English,* Sixth edition. 2000. Oxford: Oxford University Press.

OALED: *Oxford Advanced Learner's Encyclopedic Dictionary.* 1992. Oxford: Oxford University Press.

OED: *Oxford English Dictionary.* 1928. Oxford: Oxford University Press.

Sinclair, J. 1984. 'Naturalness in Language.' In *Corpus Linguistics: Recent developments in the use of computer corpora in English language research.* J. Aarts and W. Meijs, (eds.). Amsterdam: Rodopi.

Sinclair, J. 1987. *Introduction to the Collins Cobuild English Language Dictionary.* London: Collins.

Sinclair, J. 1995. *Introduction to the Collins Cobuild English Dictionary,* Second edition. London: HarperCollins.

Stubbs, M. 1995. 'Collocations and Semantic Profiles: On the cause of the trouble with quantitative studies.' In *Functions of Language,* 2, 1.

PASSPORT – Second Edition

The PASSPORT English Learner's Dictionary has undergone its first thorough editorial revision since publication in 1996. The new Second Edition now serves as the basis for new language versions appearing by the end of the year.

Created by Yaakov Levy and Raphael Gefen, PASSPORT is designed for young pre-intermediate learners, in particular pupils in junior high (lower secondary) and in the upper grades of primary school, and weaker learners in high (upper secondary) school.

The language versions published so far include Bulgarian, Chinese (Traditional), Czech, Estonian, Hebrew, Italian, Latvian, Lithuanian and Thai. In preparation are Chinese (Simplified), French, Greek, Malay, Romanian and Slovak.

The recent revision has profited and incorporated the feedback received from publishing partners in different countries, and can be divided as follows:

- corrections, changes and additions to the existing entries, with an emphasis on more synonyms and references to American/British English;
- the introduction of new entries including (a) new words widely used at all levels, especially but not only in the field of hi-tech and computers (e.g. email, e-learning, Internet), (b) existing everyday words which were not included in the earlier version but are more common now (e.g. ethnic), and (c) new meanings to existing entries (e.g. cool);
- in a number of cases, entries specifically relevant to different languages and societies have been entered at the request of the local editor.

The English-L1 section of PASSPORT has 12,000 entries and 16,000 meanings. Each version includes an extensive L1-English part, appendices on the English language and grammar, dictionary exercises, and illustrations.

The PASSPORT Electronic Dictionary is being adapted in line with the recent editorial revision and its software upgraded. The first new versions to incorporate the PASSPORT revision in electronic form will be Traditional Chinese and Greek.