Introduction:

Corpus linguistics: past and present

Antoinette Renouf
University of Central England, Birmingham

1. Introduction

I am greatly honoured to be invited by his former PhD students to contribute to this volume of papers dedicated to Professor Yang Huizhong. Unlike many other contributors to this volume, my first encounter with Professor Yang was not as his student. We met some twenty years ago at the University of Birmingham, where I was managing the Cobuild corpus-based lexicographic project (Renouf, 1987a), and Professor Yang was spending a sabbatical year. His purpose was to acquaint himself with the latest developments in corpus linguistics, and the outcome of his visit was the design of the]DEST technical corpus) (now the J i*aoda Corpus of English for Science and Technology).*

Our early discussions concerned, among other things, the validity and feasibility of taking *randomness* as a principle for *text selection* and *sampling.* Such issues<Z) were of the essence at a time when we were pushing against the limits of computing technology, and even a small corpus was still a challenge. In Cobuild, our purpose was to build as large a corpus as possible, one we felt would be adequate to support the description of *'general'* English language. Professor Yang was facing the equally daunting task, of gathering a corpus of *'technical'* languageCID**.** We naturally shared concerns about the best way to interpret these concepts through our text selection. We also grappled with the concept of corpus *representativeness;* of how to create a corpus that reflected the totality and typicality of the textual domain in question. My solution for Cobuild was to select texts according to *breadth* and *variety:* selecting a wide *range* of circumstantial parameters, associated with author, publication date, topic and so on, in the hope that this matrix of features matched the range and diversity of current language use.
Professor Yang's solution, meanwhile, was to conduct a *random* sampling exercise across the volumes in a technical library.

My contribution to this volume takes the form of a brief overview of English corpus linguistics from the 1980 to the present day, and the prospects for its application in language learning and teaching. In what I say, I am responding in part to questions

posed by Chinese colleagues.

2. What is corpus linguistics?

Corpus linguistics is by definition a branch of linguistics, the study of language. Its primary objective is to discover the facts of the language. It represents a particular approach to linguistics, one consisting of the empirical observation and analysis of authentically-occurring text, both spoken and written, as reflected in a corpus, typically electronically-held. The term' corpus linguistics' is in fact imprecise, since it is used variously to refer to every stage and aspect of activity associated with the study, from corpus design, to the practical task of corpus creation, to the actual analysis and description of the resultant data set. 'Corpus linguistics' is routinely confused with 'computational linguistics', an older term which refers primarily to the model-based, more theoretical studies that existed prior to the advent of corpora. Corpus linguistics is essentially an arts-based discipline, while computational linguistics has a mathematical heritage, and though the latter is now increasingly engaging in textual study, the approaches remains philosophically distinct.

3. Is corpus linguistics an independent subject discipline?

Some linguists ask whether corpus linguistics can develop, or has developed, into an independent *discipline.* The term 'discipline' has been defined as "a branch of learning or instruction" (Collins English Dictionary, 2001), or "an area of study, especially a subject of study in a university" (Collins Cobuild Dictionary, 1995). We can say that phonology is an area of study and a branch of learning; it has developed a terminology and methodological conventions for the study of language at the level of phoneme, and it has accumulated a body of knowledge, stored and accessible in the form of published inventories, analyses and descriptions. This applies likewise to the disciplines of syntax and morphology. It is more difficult to assign the status of discipline to corpus linguistics. Corpus linguistics does have a defined object of study, in that it requires language to be incarnate, in the form of text, and confines itself to a specified written or spoken text corpus to which it attributes theoretical validity. Like the above disciplines, it tends to accept the theoretical notion and physical reality of basic units of text such as phoneme and syntagm, as well as sub-words, words and phrases, and indeed its bread and butter involves the scrutiny of such units. It has a terminology and, let os say, an optional battery of methodological routines and strategies; it often applies quantitative measures. But it remains a paradox within the panoply of science, an amalgam of great precision and best endeavours; a somewhat undisciplined discipline. Yet as a branch of empirical study, this is ultimately its purpose, for empiricism precludes any a-priori

assumptions.

It seems to me that the crucial issue in defining the status of corpus linguistics is that, in addition to having evolved the above characteristics and concerns, it now rests on an accumulated body of research findings, knowledge and experience. Crucially, it has become self-reflective and self-critical, and in the light of the mature discussion and
debate that is now apparent at corpus linguistics symposia (c. f. Renouf & Kehoe, 2005; Renouf, forthcoming a), I am inclined to view corpus linguistics as having reached a state of maturity which one associates with a discipline.

4. Is corpus linguistics a methodology?

It is often claimed that corpus linguistics is a *methodology.* At a trivial level, it is true that it routinely involves the observation of an object of study, a word or phrase, which is typically presented in the form of KWIC (keyword in context) concordance lines, and this presentation inclines the researcher to scan the item serially within an ordered, usually alphabetical context. Even at this level, however, many other presentational formats are also possible, involving layouts not scannable in the same way. Corpus linguistics is silent on the mechanics of study: whether the eye may travel back up the page, having scanned down, and so on. Corpus linguistics also has no specified convention for matching a hypothesis against textual reality, or vice versa, or even a requirement for an articulated hypothesis at all. (We have been reminded recently by Tognini-Bonelli (2001), of the variety of methodological approaches adopted, and of the distinction existing between *'corpu:rassisted'* and *'corpu:rdriven'* linguistics~. The former licences a number of a-priori strategies, such as the annotation of corpora with grammatical tags based on intuition. *'Corpu:rdriven'* implies the iterative, bootstrapping creation and analysis of a corpus according to its internal linguistic features.
These features cannot be specified a-priori, but have to emerge from the text itself.)
Corpus linguistics furthermore does not espouse particular statistical methods, or demand statistical rigour, even though some' statistical measures (e. g. relative frequency, chi-square) are commonly applied. In short, corpus linguistics is a tool in the gift of the user, not a methodological orthodoxy.

5. Is corpus linguistics a science?

Whether corpus linguistics constitutes a 'science' in the conventional sense is debatable. It is of course unique in having itself as its object of study. It inhabits a self-reflective meta-realm, using language to articulate the study of language, as revealed within a corpus of texts. It does have much in common with the 'hard' sciences, in that it is based on a macro-theory and often on a micro-theory of language, and it is practised in

spoken language cannot be known or quantified. unless the corpus contains all the works of a given author, it can never represent the totality of language, and thus never be truly representative of it. As I have said (Renouf. 1987a), the best one can hope for is that it is sufficient and relevant for the particular research question in mind. All manner of selectional parameters are proposed to compensate for this lack: corpus creators talk of *range. coverage* and *balance* based on the circumstances of production. and (often in prefaces and sales material related to corpus-based reference materials) of *buyers* and *readership* in relation to the circumstances of reception.

9. Caveat emptor: what precautions should the corpus linguist take in working with corpora?

We have said that a corpus cannot be representative of the language. Corpus study is only as good as the corpus data is *sufficiently large, relevant* to the research question, and *authentic.*

Sufficient data is that which is large enough to allow the required features to occur abundantly. If the focus is on common linguistic phenomena such as grammatical words or high-frequency nouns and verbs. a smaller corpus may be sufficient. If the focus is on rare items. or on word combinations. only a very large corpus will sometimes suffice. unless the domain can be specified with more precision.

A delicate touch is required in using small corpora. After all. they not only do not support any statements about language that does not occur in them. they also provide weak evidence of what does. A corpus is a collection of utterances and writings by individual members of a speech community who. it is assumed, tend to observe the conventions. The smaller the corpus, the fewer the instances of each phenomenon. and the weaker the guarantee that these represent mainstream usage and are not idiolectal or erroneous. In the example below. the lexicographer found two instances of the word *earmark* used metaphorically. and having weak intuitions about ~his use. mistakenly took its presence in the corpus as endorsement of its bona fide usage. The entry appears in the first edition of the *Collins-Cobuild Dictionary* (1986) (removed from later editions)

Earmark

"If something has the earmarks of a particular thing. it has features which enable you to recognise it as being of a particular type. e. g. *it had all the earmarks of something prepared for a past college exam.* .. *this had all the earmarks of a moral dilemma".*

An earmark is in fact an identifying mark literally made in the ear of a farmed animal, such as a cow or deer. The word here should properly be either *hallmark* or simply *mark.*

On the other hand, one should guard against the assumption that all rare corpus items are erroneous, an assumption made universally in natural language processing which impoverishes many results involving indexing. The rare or single occurrence of a phenomenon is only erroneous in about 10% cases, and is more likely to be exciting new evidence of a bona fide new, reviving or stably rare item.

Relevance is defined in terms of the research question which the corpus is designed to answer, and will involve such considerations as textual domain, date of authorship and language variety. In the face of the resources required to create a sufficiently large and relevant corpus, an already available corpus can often be pressed into service to represent far more than it was designed for.

A large relevant corpus still only provides partial evidence of a linguistic feature, and it does not, as said earlier, support statements about language not occurring in it.

The very frequent occurrence of a feature probably reflects its importance and versatility in the language, though it might also mean that the corpus is skewed towards a particular genre. For instance, a corpus of newspaper text can be regarded as a 'general' corpus for some purposes, but as having some features peculiar to a particular domain or social register, for others.

Another crucial issue is *authenticity.* Authentic text is that which is produced with a straightforward communicative purpose and which can be assumed to exemplify mainstream usage. It is to be distinguished from *poetic* or *literary* text, which by definition flouts mainstream usage to stylistic effect; also from *concocted dialogue* in novels and drama, where words are being put into the mouths of speakers; as well as *simplified* text, such as is found in readers for learners of English; and more or less straightforward *langue,* as found in pedagogic course books (e. g. Ramer, in Renouf and Kehoe, 2005). There was a debate in the early 80s over the issue of *naturalness* (Sinclair, 1985), which was defined as a quality perceived immediately as lacking by native speakers in encountering inauthentic data.

In addition to the necessity for the *data* to be adequate, the corpus *users* must also be in a position to conduct and draw sensible conclusions from the results of their search. Thus, all corpus users ideally require training in taking *appropriate corpus based approaches,* in selecting *appropriate corpora,* and in the sensible use of *corpus-analytical tools.*

They need to be aware of the status and the layout (including the principles underlying any annotation) of their data. They need to appreciate the nature of the

structure of text: that a corpus will yield very different amounts of data for the commoner lexical and grammatical features of the language than for the less common (Zipf's law, 1949, of scaled probability); that the lexicon is structured with decreasing frequency into bands of words which play different roles in the composition of text, from the universally frequent lexical words, to the grammatical words, discourse organising words, stylistic markers, and technical or topic words (Renouf, forthcoming b); that common phenomena are time-consuming to study but also likely to be more reliable; that rare occurrences should be treated with healthy scepticism, and so on.

With regard to selecting corpus and corpus topic, users need to understand both what a corpus represents and what their own limitations are in knowing how to interpret it. Some corpora, for example general every-day prose such as journalism, or small, technical corpora in their subject of specialism, are more accessible to the non native speaker of English. Large corpora containing a mix of informative and literary language will pose more problems of interpretation; and the Web as Corpus is an extreme case in point**.**

With regard to the choice of corpus-based study, basic grammatical analysis and the study of collocation are within the grasp of the language learner. Other topics, where word play and figurative use of language predominate, as with idiom and metaphor, or where there are no referential or dictionary aids, as with neologistic use, require of the user a highly-developed intuition for interpreting corpus results and differentiating language error and routine rule-application from creativity and coinage.

10. What are the main developmental trends in corpus construction and corpus-based linguistic research?

I should say that there have been four major developmental trends in corpus design and construction over the last 25 years. The first has been the growth from the small to the large corpus, the second the shift from synchronic to diachronic corpora and study, the third from the designed corpus to 'large amounts of text', and the fourth from single focus to multi-dimensional corpora. Each of these stages can be directly or indirectly attributed to advances in technology, but each at the same time reflects an advance in theoretical thinking and understanding, and the level of maturity that corpus linguistics has reached as a fledgling discipline. Each has had an impact on the evolution of corpus-based linguistic research.

10.1 Small corpus to super-corpus

Prior to the 70s, there had been small-scale attempts to conduct empirical. quantitative, word-based studies on such small corpora as computing technology would allow, of which Sinc1air et aI's 1970 'English Collocation Studies: the OSTI Report'

(Krishnamurthy, ed. , 2004) is a pioneering example. In the 70s, the one-million-word 'standard' US 'Brown' Corpus, compiled of texts from 1961 at Brown University (Francis and Kucera, 1964), and the equivalent UK 'LOB' corpus, compiled by researchers in Lancaster, Oslo and Bergen, heralded the start of corpus linguistics as we know it today. They have remained models for the small, designed corpus, including the Kolhapur Corpus (Indian English), the first version of the JDEST Corpus, and many other, chiefly specialised, corpora. The type of linguistic research which all these early small corpora were designed and best qualified to support was grammatical. A major leap in corpus construction occurred in the 80s, when the Cobuild project, the first joint academic and industrial venture in UK arts, managed to push computing technology to its limits and establish a corpus of almost 10 million words of writing and speech.

Suddenly, lexis became sufficiently accessible as an object of study to allow the lexicographic description of English based on corpus-linguistic principles to emerge alongside grammatical description. From the mid 80s, computing technology has evolved steadily, leading to corpora in the 100s of millions of words, such as the British National Corpus and the Bank of English. One ironic effect on corpus linguistic study has been to require specialised tools and statistical techniques, such as sampling, to enable researchers to cope with the amount of data available for the more common linguistic phenomena. The more recent move from mainframe computing to PC-based activity has encouraged the design of individualised tools such as Wordsmith, devised by Mike Scott (2004), based on earlier work by him and Tim Johns, for the individual learner and teacher.

10.2 Synchronic to diachronic study

The notion that language is a changing phenomenon is long established, notably among language historians, but it was not until the early nineties that it became a focus of explicit study in corpus linguistics@. The LOB and Brown 'standard' corpora of data from 1961 were matched in 1991-6 at the University of Freiburg (Mair, 1997) by two small parallel corpora, FLaB and Frown, composed of text produced some 30 years later. This allowed those aspects of modern English language innovation and change to be identified which had clearly emerged after the 30-year gap. By the nineties, too, technology and the requisite electronic data resources were available to allow corpora to be treated as a flow of data, as open-ended entities (Renouf, 2000). The RDUES unit accumulated unbroken news data from the late 1980s onwards, thereby initiating a second approach to modern diachronic study, namely the study of short-term or 'brachychronic' change. Where this differed from the FLOB and Frown approach was in its ability to chart the rise and fall of lexical and grammatical phenomena, and particularly of new coinages, across time, though of course it was still unsure of

capturing the actual moment of creation.

10.3 Designed corpus to text collection

The early days of English corpus linguistics had seen linguists laboriously selecting,

sampling, correcting and carefully honing a small corpus so that it could be exhaustively Oohansson, 1982) studied and all phenomena within it quantified in relative terms. The time and care put in was prohibitively expensive, but was nevertheless continued with the larger designed UK corpora.

Meanwhile, in the US of the 90s, the mainstream computational linguistic community of mathematicians, AI experts, cognitive scientists and engineers, funded by the US defence agencies and industry, began to move away from Chomsky and his anticorpus doctrine as they realised that .access to real textual data might provide wealth if knowledge about language use could be automated and applied in IT contexts, including translation, language generation and knowledge management. Thus began a track in US linguistics which ran parallel but separate to the pioneering corpus linguistic work of Francis and Kucera, and the small but growing pockets of corpus linguistic activity in Arizona (Biber), Southern California (Chafe, Dubois) and Boston (Meyer). This new track concerned not the laborious design of 'balanced' corpora by linguists, but the swift accumulation of large but fairly random data collections, such as Hansard proceedings. Clearly, this type of collection eliminated the possibility of exhaustive or quantified study of corpus as honed artefact, and instead provided a rough and ready basis for hypothesis testing and inference drawing.

The advent of the World Wide Web in the mid 90s provided a further opportunity for the acceleration of the term ‘corpus’. Corpus linguists who study current language use require a large, up-to-date data source. Corpora are time-consuming and expensive to create, and most existing corpora of modern English are thus too small to support large-scale studies, out-of-date by the time they are available for use, and synchronically organised, so unable to support the study of language change. In these circumstances, *web-based text* represents a potentially valuable source of language data to supplement conventional text corpora . For the learner of English, it can be a rich source of colloquial, neologistic and rare language use. Web-based text is unconventional, but it is copious, up-to-date, updatable and principally freely available. In practice, however, considerable investment in linguistic research and software tool development is required to overcome the chaos, heterogeneity and unorthodoxy of web text and to produce satisfactory results in terms of usability and speed. Linguists have been using search engine user-front-ends, primarily ‘Google’, to gain access to such instances of language use, but these are not designed to support linguistic search, in particular pattern matching. The *WebCorp*

project (Renouf et ai, 2005) has spent the last 4 years or so dealing with the issues which arise from treating the web as a corpus, and has developed a system that can produce tailored, language-specific, dated results to help linguistic and applied linguistic researchers, translators, teachers and learners. Again, instances of particular language use extracted on-line from the web can only be considered a corpus in metaphorical terms. Only when pages and citations are down loaded and processed off-line according to a set of corpus design principles will they be amenable to traditional, exhaustive inspection and to statistical study beyond simple frequency.

10. 4 Single to multi-dimensional corpus

More recently, as the benefits of corpus study and its potential for further exploitation have been appreciated, and technology has advanced, the established range of 'general' text corpora and even of the traditional' special-purpose' corpora has been supplemented

by a new generation of multi-purpose corpora. These allow the study of language and the annotation of corpora from two or more points of view, such as *regional* and *historical variation*, (see the 'Variation and Change' projects, notably at Helsinki University and Freiburg), *multi-dimensionality, multi-linguality* in parallel corpora and *multi-media* corpora. The' *multi-layered*' corpus is one such enterprise, where the findings of the experts in different disciplines can be integrated into the meta-text in the form of cross disciplinary annotation (Meurman-Solin, in Renouf and Kehoe, 2005) which leads to new methodologies, new discoveries and indeed to new fields, or possibly sub-fields, of study (such as 'historical socia-pragmatics', Nevala, 2004).

11. Applied corpus linguistics: What does corpus linguistics offer non-native speaking researchers, language learners and teachers?

As a language researcher, I am particularly aware of how linguistic insights derived from corpus study can be applied to the field of IT, and in particular to knowledge management and the extraction of information from large electronic document databases;

as well as to human and machine summarisation and translation. But of course applied corpus linguistics also plays a key role in language teaching and learning. A corpus uniquely provides a model of real language use, whether prestigious or colloquial, typical or marginal, general or technical, depending on corpus type. From this knowledge base can be derived all manner of teaching and testing methods and material.

The applied corpus linguistic approach is exemplified by the achievements of the contributors to this' volume. Studies of the CLEC and COLSEC corpora of learner English, contrasted where appropriate with native-speaking norms as reflected in the JIAODA corpus and other sources, have resulted in descriptions of many salient aspects of Chinese learner English, which in turn represent crucial knowledge for the

development of innovative and world-leading language tests for Chinese college learners. Such work also advances the field of corpus-based interlanguage study, which was introduced in Europe through the ICLE (International Corpora of Learner English) project (Granger, 1994).

In the future, applied corpus linguistics will continue to modify its view of the nature of real language in the light of greater understanding: for instance, the full extent of the pre-fabricated, phraseological nature of text has only emerged very recently, as has the full significance of textual domain specificity. Insights from neighbouring fields such as psycholinguistics will increasingly filter in, and cross-disciplinary collaboration will increase, to the benefit of all. There will be a growing awareness of language as a changing phenomenon, and a corresponding emphasis on updating basic data sources.

Overall, prospects for the future of applied corpus linguistics in China and elsewhere are rosy. A viable infrastructure of technology, data collections and basic methodologies is in place, as a platform on which to build in the future.

References

Baker,Mona, Francis Gill and Elena Tognini-Bonelli (eds.) (1993): *Text and Technology: In honour of*] *ohn Sinclair.* Amsterdam/Atlanta, GA: John Benjamins Publishers.

Chomsky, Noam (1957). *Syntactic Structures.* The Hague, p.159.

CollinirCobuild English Dictionary (1995). Glasgow/London: HarperCollins Publishers.

Collins English Dictionary, 2001. Glasgow/London: HarperCollins Publishers.

Corder, S. P. (1974). Error Analysis. In J. P. B. Alien and S. Pit Corder (eds.) *Techniques in Applied Linguistics (The Edinburgh Course in Applied Linguistics:* 3), London: Oxford University Press (Language and Language Learning), pp. 122 - 154.

Facchinetti, Roberta (00.) (forthcoming): *Corpus Linguistics Twenty-five Years On. Selected Papers of the Twenty- fifth International Conference on English Language Research on Computerised Corpora.* Amsterdam &. Atlanta: Rodopi.

Fillmore, Charles J. (1992). 'Corpus Linguistics' or 'Computer-aided Armchair Linguistics', In Svartvik, Jan. (ed.) *Directions in Corpus Linguistics.* Berlin/New York, 35.

Francis, W., and Kucera, H., (1964). *Brown Corpus Manual.* Revised 1979. Brown University, Rhode Island.

Granger, Sylviane (1994) 'The Learner Corpus: A Revolution in Applied Linguistics', *English Today* 39, 25 - 29. Halliday, Michael. A K. (1985). *Spoken and Written Language.* Oxford: Oxford University Press. Johansson, Stig (ed.) (1982). *Computer Corpora in English Language Research.* Norwegian Computing Centre for the Humanities, Bergen.

Johns, Tim. (1988). 'Whence and Whither Classroom Concordancing?' In Bongaerts, Theo et al.

(eds.). *Computer Applications in Language Learning.* Foris.

Johns, Tim. (991). 'From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning'. In Johns and King (eds.), pp. 27 - 45.

Johns, Tim, and Philip King (eds.) (1991). *Classroom Concordancing, ELR journal* 4. University of Binningham.

Krishnamurthy, Ramesh (ed.) (2004). *English Collocation Studies: The OSTI Report,* by Sinclair, John McH, Jones, Susan and Robert Daley. Continuum Books, London and New York. Ljung, Magnus, (ed) (997). *Corpu~Based Studies in English.* Amsterdam & Atlanta: Rodopi. Mair, Christian. (997) 'Parallel Corpora: a Real-Time Approach to the Study of Language Change in Progress'. In: M Ljung, M (ed.) (997). 195 - 209.

Meunnan-Solin, Annelie (forthcoming). 'The Manuscript-Based Diachronic Corpus of Scottish Correspondence' in Joan Beal, Karen Corrigan and Hermann Moisl (eds), *Models and Methods in the Handling of Unconventional Digital Corpora* . Volume 2: Diachronic corpora. New York: Palgrave Macmillan.

Nevala, Minna (2004). 'Address in Early English Correspondence. Its Forms and Socio-Pragmatic Functions'. In *Mbrwires de la Societe Neophilologique de Helsinki* 64. Helsinki: Societe Neophilologique.

Renouf, Antoinette (1987a). 'Corpus Development', in Sinclair, John McH (ed.) *Looking Up,* Glasgow/London: HarperCollins Publishers, pp. 1 - 40.

Renouf, Antoinette 0987b): 'Moving On', in Sinclair, John McH (ed.) *Looking Up,* Glasgow/ London: HarperCollins Publishers, pp. 167 - 178.

Renouf, Antoinette (2000). 'The Time Dimension in Modern English Corpus Linguistics', in Kettemann, Bernhard and Georg Marko (eds.), *Language and Computers, Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz* 19 - 24 *j uly, 2000.* pp. 27 - 41(5).

Renouf, Antoinette and Andrew Kehoe, (eds) (2005). *The Changing F ace of Corpus Linguistics.* Amsterdam & Atlanta: Rodopi.

Renouf, Antoinette, Kehoe, Andrew and Jayeeta Banerjee (2005). 'The WebCorp Search Engine: A holistic approach to web text search', in *Electronic Proceedings of CL2005,* University of Binningham.

Renouf, Antoinette (forthcoming a). 'Corpus Development 25 years on: from supercorpus to cyber corpus', in Facchinetti, Roberta (forthcoming). ~Renouf, Antoinette (forthcoming b). 'The Lexical Structure of Text' (provisional title). . Ramer, Dte (2005). *'Looking at Looking: Functions and Contexts of Progressives in Spoken English and' School' English'*, in Renouf and Kehoe, 2005.

Seou, Mike (2004) *WordSmith Tools* version 4, Oxford: Oxford University Press. ISBN: 0 - 19 459400 - 9. http://www.lexically.net/wordsmith/

de Saussure, Ferdinand (915). *Cours de linguistique generale.*

Sinclair, John McH. (1984). 'Naturalness in language use', in: *Lexisand Lexicography.* Singapore: National University Press, 96 - 104.

Sinclair, John McH (ed.) (1986). *Collin~Cobuild Language Dictionary,* Glasgow/London:

HarperCollins Publishers.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work.* Studies in Corpus Linguistics 6. Amsterdam/ Atlanta, GA: John Benjamins Publishers.

Zipf, George. (1949). *Human Behavior and the PrinciPle of Least Effort.* Addison-Wesley, Cambridge MA

specified. However, this is a situation which, as said, is shortly to be radically improved upon.

@ The work of historical corpus linguists was diachronic in the global sense that between them, historical linguists studied language at different stages, though the individual studies could be synchronic.

@ It is freely available to Chinese users, via our web-site, at http://www. webcorp. org. uk. Nonacademics such as professionals and business-people, who have received a formal education but had little or no training in writing, can also use the *WebCorp tool* as a usage guide, as can members of the public.

@ The purpose was to record and analyse the typical interlingual features of a learner's particular speech group, and compare these with the equivalents in a more standard native-speaking corpus, in order to help the learner to evolve an analytical awareness of his/her own language practice, and to modify this to approximate native-speaking norms. A subsequent purpose, now widely in progress, is to compare the interlanguages of many different languages, in order to identify features which common to many languages.