Corpus development 25 years on: from super-corpus to cyber-corpus

Antoinette Renouf

formerly the University of Liverpool, UK
(now the University of Central England, Birmingham)

Abstract

By the early 1980s, corpus linguists were still considered maverick and were still pushing at the boundaries of language-processing technology, but a culture was slowly bootstrapping itself into place, as successive research results (e.g. Collins-Cobuild Dictionary*) encouraged the sense that empirical data analysis was a* sine qua non *for linguists, and a terminology of corpus linguistics was emerging that allowed ideas to take form. This paper reviews the evolution of text corpora over the period 1980 to the present day, focussing on three milestones as a means of illustrating changing definitions of 'corpus' as well as some contemporary theoretical and methodological issues. The first milestone is the 20-million-word* Birmingham Corpus *(1980-1986), the second is the 'dynamic' corpus (1990-2004); the third is the 'Web as corpus' (1998-2004).*

1. Introduction

I have been invited to review developments in corpus linguistics over the last 25 years, up to the present day. In fact, given the immensity of the topic, I have decided to focus on corpus creation rather than on corpus exploitation. I shall take the briefest of glances into the future. I am honoured to follow on as speaker from Jan Svartvik, a pioneer in the field in whose corpus-building footsteps I followed only in 1980.

I am no longer the young thing I was when I attended my first ICAME conference in Stockholm in 1982, but I can at least report that I entered corpus linguistics after the first generation of terminology, such as *mechanical recording*, which is such a giveaway to one's age. I was of the generation of UK corpus linguists who by the early 1980s had established that the plural of *corpus* was probably *corpora*, and who, were behind closed doors, decadently using *text* as a mass noun, *data* as a singular noun, and the term *concordances* to mean 'lines within a concordance'.

In this paper, I shall give a brief overview of the history of corpus development, starting with a backward perspective echoing some of Jan's talk, but focussing on the more recent events: the larger 'super-corpora' of the 1980s and 1990s, and the

current and future 'cyber-corpora'. I shall refer mainly to the text collections for which I have personally been responsible and am thus better equipped to comment on: the 18 million-word *Birmingham Corpus* (which evolved into the Bank of English); the open-ended corpora of present-day journalism which my Unit has been processing chronologically over a period of 15 years so far; and most recently, the 'Web as corpus', the ad-hoc, always changing corpus of language data extracted from web-based texts. In the course of this review, I shall propose a model to explain the particular path that corpus development has taken, and discuss the theoretical and practical issues involved at different stages of its evolution.

The phases of corpus evolution approximately follow the pattern in Figure 1. The dates there are presented as starting dates since they each refer not just to specific corpora, but to *types*, *styles* and *designs* of corpora which continue to be constructed into the 21st century. The small corpora *LOB* and *Brown* have been followed by their updated equivalents, *FLOB* and *Frown*;[1] the super-sized *British National Corpus* (*BNC*)[2] emerged in the wake of the *Bank of English*; the small, specialised *MICASE Corpus of Academic Speech* built on earlier work in spoken corpus production, and so on.

⇓ 1960s onwards: the one-million word (or less) Small Corpus
- standard
- general and specialised
- sampled
- multi-modal, multi-dimensional

⇓ 1980s onwards: the multi-million word Large Corpus
- standard
- general and specialised
- sampled
- multi-modal, multi-dimensional

⇓ 1990s onwards: the 'Modern Diachronic' Corpus
- dynamic, open-ended, chronological data flow

⇓ 1998 onwards: the Web as corpus
- web texts as source of linguistic information

⇓ 2005 onwards:
- the Grid; pathway to distributed corpora
- consolidation of existing corpus types

Figure 1: Stages in English language corpus evolution

2. Major drivers in corpus development

My conception of corpus development is that it has been shaped by three major *drivers*, or motivating forces, over the years. Of course, this is a stylised, simplified representation of what were typically complex and many-layered circumstances and decisions. These drivers are what I shall characterise for convenience as *science* (or intellectual curiosity), *pragmatics* (or necessity), and *serendipity* (or chance).

The first driver, *science*, is the proper motivation for academics and thinkers. By *science*, I mean the desire to undertake an empirically-based methodological cycle, beginning with curiosity based on introspection, intuition and probably data observation, which leads to the formulation of a research question or hypothesis or conviction, which in turn leads to a programme of hypothesis-testing through data observation, which leads to a discovery, which provides knowledge and experience, which in turn fosters curiosity and the development of further hypotheses to explore.

The second driver in corpus development, *pragmatics*, or necessity, is less exalted. On the one hand, given the time and expense involved in corpus creation, the design decisions taken are inevitably influenced by the availability of particular data, technologies and funds. On the other hand, corpus creators are subject to the vagaries of external agencies, ranging from publishers' requirements, to priorities set by funding bodies, to the exigencies of governmental research grading.

The third driver for corpus development, *serendipity*, is also a powerful one. Jan Svartvik (this volume) reminded us in the context of his creation of the *London-Lund Corpus* that once a concept is demonstrated to be possible, it does not take long for it to be taken up by others. Equally, new data or technology or other innovation can emerge at any moment to support hitherto impossible corpus initiatives. Or vital resources may unforeseeably become available.

Applying the three-category model to the small corpora of the 1960s exemplified above, I would say that the primary driver was *scientific*, and the theoretical underpinnings were as follows:

- that language in use is a bona fide object of study;
- that 1 million words was sufficient to ensure adequacy of grammatical description;
- that exhaustive study was the right approach to a body of text;
- that a corpus could *represent* the language.

More recent small corpora have obviously been built with the benefit of hindsight. My analysis of the continuing creation of small corpora when it is

technologically possible to create larger ones is that here necessity is playing a larger role. Many newer small corpora are designed to be comparable with earlier models (e.g. *FLOB* and *Frown*; perhaps learner corpora), some are small because they are very specialised, others are relatively small because of the scarcity of data; e.g. *ZEN* (Fries 1993); the *Corpus of Early English Correspondence* (Nevalainen and Raumolin-Brunberg 1996 and other historical data collections), and many others still are constrained by limited funds and the time pressure to produce results quickly (spoken corpora being particularly costly).

3. The evolution of the super-corpus: 1980s onwards

3.1 Examples of 'super-corpora'

Birmingham Corpus (1980-1986)
University of Birmingham
20 million words British/American English, written/spoken text
Bank of English (1980-)
University of Birmingham
500 million words British/American English, written/spoken text
British National Corpus (1991-1995)
Longman, Oxford University Press, Chambers
University of Lancaster
Oxford University Computing Services
100 million words British English, written/spoken text

3.2 Drivers for the super-corpus

The same *scientific* curiosity which led to the creation of the small, standard corpora underpinned the development of the first of the next generation of 'super-corpora', the *Birmingham Corpus* of 1980-1986. There was continuing curiosity about the nature of real language use, and a desire to discover further unknown facts of the language through as exhaustive study as possible. The difference was that, by the 1980s, there was a realisation that there were questions about lexis and collocation, and indeed even about grammar, which could not be answered within the scope of the small corpus. In addition to scientific curiosity, serendipity assisted the sea change in corpus linguistic thinking. With the emergence of the first corpus-based lexicographic products, perceptions changed within the major publishing houses, and suddenly it became desirable and even indispensable to pay at least lip service to the virtues of corpus-based lexicology. Equally fortuitously, developments in computing technology were creating an ever less hostile climate, providing hitherto undreamed of opportunities for such research.

3.3 Practical issues in the early days of creation of the super-corpus

In the 1980s, whilst the conditions became possible for larger-scale corpus development, there were still formidable practical issues to be overcome. Within the Collins-Cobuild project at the University of Birmingham, Jeremy Clear and I, for instance, had to create the *Birmingham Corpus*, a collection of as near 20 million words as possible, in one year, so that Collins-Cobuild lexicographic work could meet its deadlines. Dashing around like mad things, we were the Ginger (given Jeremy's colouring) and Fred of corpus building. To begin with, the conversion of printed text to machine-readable form was in its infancy and there were two options for this initial process. The first was to use the so-called Kurzweil Data Entry Machine (KDEM) (Kurzweil 1990), an early optical scanner the size of a large photocopier, ingeniously designed to convert printed text into Braille. The deciphering capability of this machine was limited in comparison to that of today's scanners, and it tended to offer the character 'W' as a candidate for any character or blob (wryly dubbed the "is it a W?") response, fairly regularly whenever book pages, particularly paperbacks, contained less than perfect print quality. We had two operators working simultaneously non-stop for many months, and to keep up production I had to acquire special dispensation to have women students as well as male working overnight on campus; I processed many books myself. There was a point of diminishing return at which the theoretically slower option, keying-in, became more efficient. This was regularly reached with thinner newspaper pages, through which the KDEM shone a light and diligently tried to read text on both sides simultaneously.

Another practical issue was the acquisition of the texts themselves, and in particular the 'right' edition. The KDEM required two copies of each book, one of which had to be dismembered for scanning, the other required for checking. Books quickly go out of print, and finding two copies of each, dating back to the 1960s, was a headache involving the continual scouring by me of the groves of Birmingham second-handia. The concomitant task of acquiring permission to reproduce the content of hardcopy books in electronic form was a procedure I endured over a 5-year period. The rights manager in most publishing houses was actually a series of individuals who radiated all the joy and urgency of being on fatigue duty. Copyright can be held separately, often by different parties, for the UK, Europe, US and North America, Bermuda, the Cayman Islands, and any combination of these and other remote islands and protectorates. Furthermore, books that are no longer best-sellers and thus present no copyright problems in principle can suddenly be serialised for TV, become overnight successes, and renew their copyright status. In the early days, having to extract these data retrospectively from concordanced output would have caused major logistical problems. Fortunately, only one writer, J. D. Salinger, refused permission for his classic novel *Catcher in the Rye* (1951), unambiguously and in advance, and just a couple of the hundreds of newspaper articles required modest payment for use.

The next issue, text processing, had to be tackled by Jeremy Clear. This was a struggle in the early 1980s. Jeremy progressed from using the punch cards mentioned by Jan Svartvik (this volume) to overnight batch jobs which regularly failed and had to be resubmitted or which, if he was lucky, generated vanloads of output on streams of perforated A3 paper. Jeremy had to contend with the limited processing capacity that was available on early mainframes, of which the first, a UK-manufactured ICL 1906A, was the proverbial walk-in store-room which Jan Svartvik described. According to Jan, Henry Kučera reported that the concordancing of the one-million-word *Brown Corpus* took the total mainframe capacity of Brown University Computer Unit for a day. To process the first-stage 7.3 million-word *Birmingham Corpus*, Jeremy had to commandeer the entire University mainframe resources, and process the data, in 1.2 million word chunks, into 6 batches of concordances, over eight successive weekends. Once he had processed the text, there was still the problem of limited on-screen access to concordances. In the first year, lexicographers had to work across six sets of microfiches, each alphabetically-ordered. That is, to analyse the word *apple*, they had to look in six places. Of course, there were no PCs, and in-team communications were still paper-bound.

Data storage and processing quickly became easier by 1983, and we were able to move with it, to a larger and more manipulable corpus. Nevertheless, in 1983, the necessity for data compression over larger-scale storage was briefly contemplated.

3.4 Theoretical issues concerning the large 'general' corpus

The first issue for the *Birmingham Corpus*, as for the earlier 'standard' corpora, was theoretical: how to create a body of text which could be claimed to be an authoritative object of study. Ideally, it would be *representative* of 'the language as a whole'. But given the unknowability of that totality, and hence the impossibility of designing a perfect microcosm, we justified our design strategies in relation to a network of more readily accessible criteria, such as:

linguistic parameters

- fiction versus non-fiction
- speech versus written text
- authenticity[3]
- regional and social varieties, domain specificity; generality
- text of particular era and spanning specific time-span statistical parameters
- breadth, variety, 'principled selection', relevance and sufficiency
- balance
- sampling, of text extracts versus whole documents; and applying 'random' versus 'non-random' rules of selection

demographic parameters

- age, gender, social grouping and ethnicity of author

- research needs of corpus users

Though the *Birmingham Corpus* did not exploit them, additional demographic selectional criteria were also available to commercial publishers, such as:

- focus on receiver rather than producer of language
- readership and other sales figures)

Through the 1980s and early 1990s, though it was generally accepted among corpus creators that *representativeness* was unattainable, it was felt necessary to present selectional criteria in those terms. Extracts from some early corpus design rationales are quoted here. Looking back to a chapter in *Looking up* (Sinclair 1987), the companion guide to the *Collins Cobuild Dictionary*, I see that I wrote:

> [The *Birmingham Corpus* was] designed to *represent* the English language as it was relevant to the needs of learners, teachers and other users, while also being of value to researchers in contemporary English language. (Renouf 1987: 2)

while Della Summers (1993) said of the *Longman/Lancaster English Language Corpus* that it was:

> *representative* of the standard language in a very general sense, not restricted to a regional variety (e.g. British English or a local dialect) or a narrow range of text types (e.g. scientific texts, newspaper writing, language of a particular social class). (Summers 1993: 186)

Concerning the *BNC*, text selection was characterised in the following terms:

> text that is published in the form of books, magazines, etc., is not *representative* of the totality of written language that is produced. […] However, it is much more representative of written language that is received, and… thus forms the greater part of the written component of the corpus. (Burnard 2000: 1)

Meanwhile, Matti Rissanen (1992) acknowledged, of the *Helsinki Corpus of English* that:

> Just as a corpus will never reliably reflect the language in all its varieties and modes of existence, so, too, parameter coding can never hope to give a complete and theoretically valid description of the samples. (Rissanen 1992: 188)

As the *BNC* was emerging, the achievability of true *representativeness* was historically debated by John Sinclair and Willem Meijs at a 1991 conference in

St. Catherine's College, Oxford. This grail continues to haunt corpus linguists as we progress into the new millennium. See, for example, even as recently as 2004, the stated ambitions for the new *Corpus of Spoken Israeli Hebrew* (CoSIH) at Tel Aviv University (Izre'el and Rahav 2004) are as follows:

> Our aim is to produce a *representative* sample which will take into account not only demographic criteria but also account for contextual varieties. Thus, data should be collected according to two distinct types of criteria: while sampling is conducted according to statistical measures and thus will be *quantitatively representative* of the entire population, collecting data according to analytical criteria is not necessarily *statistically representative*; the *representativeness* of a corpus must be based on the internal social structure of the speech community for which it is designed. (Izre'el and Rahav 2004: 7)

Stepping back, the only progress that I have noticed in this debate over the last twenty-five years, which is presented each time as if it were a new problem, is a change in the terminology used to conduct it! According to our *WebCorp* output, the variant form *representativity* is now in vogue:

1. There is now a debate among corpora research community concerning this concept of the **representativity** of the data gathered inside text corpora
2. So it seems that we now witness a move from **representativity** towards reusability
3. How can we measure the **representativity** of a corpus with respect to a given linguistic construct?
4. That raises the problem of the **representativity** of the data base and of the application of methods for the presentation of findings.
5. We consider issues such as **representativity** and sampling (urban–rural, dialects, gender, social class and activities
6. In search of **representativity** in specialised corpora
7. The discussion of issues of corpus annotation, the **representativity** of corpora, economy, and an optimized structuring of the data
8. Their **representativity** is measured by reference to external selection criteria
9. what is meant by 'spoken' and 'written', the **representativity** of a certain type of language
10. The twin theoretical problems of data **representativity** and corpus structuration.
11. the difficulty of defining **representativity** in such a corpus.

Figure 2: *WebCorp* output for the term *representativity*, May 2004

A fundamental tenet of corpus linguistics in the 1980s (Johansson 1982) was breached by the larger corpus, which was that a corpus should be studied exhaustively. The purpose had been to ensure that the linguist exploited a small resource exhaustively, so as not to miss any aspects of language use of which

he/she had not been aware, and with a view to applying quantitative measures to establish the relative significance of phenomena in the corpus. The small corpus was a precious, hard-won object. The advent of larger corpora necessitated new analytical approaches. While some phenomena remained sparse even in the huge text collections, many more words were well represented, and some of the commonest phenomena in text, the grammatical and phraseologically productive lexical items, now occurred in such vast numbers that they could not always be studied exhaustively, but had to be sampled (as in the Cobuild lexicographic project, for instance).

3.5 Was building the large *Birmingham Corpus* worth the effort?

Yes, it was. It revealed the nature of lexis and collocation, and underpinned the ground-breaking *Collins-Cobuild Dictionary* as well as countless other books and theses then and since. Importantly for subsequent applied research, it also revealed the relationship between surface patterns of text and meaning. The paradigmatic axis of language only being realisable syntagmatically in text, the link between collocation and word meaning is readily identifiable at the surface level. In the course of time, the awareness of this fact allowed me to devise and set up some large-scale projects – AVIATOR[4] (Birmingham 1990-1993, ACRONYM[5] and APRIL[6] (Liverpool 1994-2000) - in which, in collaboration with a series of inventive software engineers, including Alex Collier, Mike Pacey and now Andrew Kehoe and Jay Banerjee, I have been able to exploit the regularity of word collocation in text to the limits. In AVIATOR, for instance, the change in meaning of a word was identified automatically by a significant change in its collocational patterning in a corpus of journalism. In ACRONYM, by dint of the simple comparison of the collocational profiles of words, we can find synonyms and near synonyms ('nyms') automatically. In 2004, ACRONYM produced the results, shown in Figure 3 for the word *solace*.

comfort	excuse
consolation	happiness
inspiration	cure
satisfaction	counselling
encouragement	respite
reassurance	salvation
warmth	succour
refuge	sanctuary
shelter	remedy
punters[7]	haven

Figure 3: ACRONYM extract of ranked 'nymic' output for the term *solace*, *Independent* news text, 2004

Moreover, in 1996 for the word *surgery*, it produced the multi-word 'nyms' shown in Figure 4.

heart transplant	coronary event
heart surgery	median survival
heart transplantation	surgical procedure
reduction surgery	surgical resection
breast biopsy	surgical intervention
coronary angioplasty	artery bypass
coronary angiography	outpatient surgery
coronary revascularization	prophylactic mastectomy
coronary stenting	angioplasty procedures

Figure 4: ACRONYM extract of ranked multi-word 'nymic' output for the term *surgery*, *Independent* news text, 1996

These were clearly promising first-stage results, unedited as they are, in themselves raising important questions about the nature of lexical semantics in use, and having enormous potential for applications in fields from lexicography to database management.

4. The evolution of the modern diachronic corpus

The concept of a diachronic, or 'monitor', corpus had been raised as a theoretical possibility back in 1982 by Sinclair (Johansson, 1982). His vision had been of language as a changing window of text, in which the computer would 'monitor' aspects of language use across time, but would then discard each stretch of text once processed.

It was not until 1990 that the first 'dynamic' corpus of unbroken chronological text was finally established by the RDUELS Unit in our 1990-1993 AVIATOR project at Birmingham, using text from the *Times* newspaper dating back to 1988. By that time, there was no obstacle to its being archived in its entirety, thus allowing the flexibility to return to the data as new hypotheses emerged for examination. The second 'dynamic' corpus of this kind was set up by the RDUES Unit in the ACRONYM project at Liverpool in 1994, this time with *Independent* news text, also starting from 1988, the date of the inception of that newspaper.

There are three types of corpus which currently support diachronic study to the present day, each representing different approaches to diachronic study. The first type comprises the small, synchronic but parallel 'standard' corpora, *Brown*, *Frown*, *LOB* and *FLOB* (Brown University, Universities of Lancaster, Oslo, Bergen and Freiburg); the second type is the chronologically-ordered corpus of text samples of historical English register now reaching into the 20th century, known as the *Archer Corpus* (Universities of Arizona, Southern California, Uppsala and Freiburg). The third type is represented by the afore-mentioned unbroken, chronological data flow of *Times*, and more recently of *Independent*

and *Guardian* journalistic text (now at the University of Central England, Birmingham).

4.1 Drivers for the modern diachronic corpus

The motivation for the development of the modern diachronic corpus was primarily *scientific,* or theoretical. It manifested itself in the following ways:

- awareness that language is a changing phenomenon;
- belief that language change can in principle be observed in corpus data;
- curiosity about innovation, variation and change in grammar and lexis.

In our case, it involved a curiosity about neologistic assimilation, lexical productivity and creativity; and the structure of the lexicon at the level of hapax legomenon. I was also curious to investigate the power of collocation to identify change in word use, in word sense and in meaning relationships between words. Supporting this theoretical impetus for the modern diachronic corpus was the serendipitous availability of news text held as an electronic flow, and of the necessary funding.

4.2 Theoretical issues concerning the 'dynamic' diachronic corpus

There are fundamental differences between static and dynamic corpora, both in purpose, design and in methodology. The purpose of a dynamic corpus is to support the study of language change over time. Monitoring chronologically-held text reveals innovation, trends and vogues, revivals, patterns of productivity and creativity. In return, however, the goal of quantification and thence the assignment of significance which was taken for granted with the static, finite corpus becomes impracticable with the open-ended flow of text, since it requires knowledge of the size of the total population.

Different theoretical approaches are taken to the study of modern diachronic corpus data. The clearest distinction lies on the one hand between the study of language change across a significant period in two parallel finite corpora, such as *Brown* and *Frown*; and on the other, the ongoing study of change in an open-ended, unbroken flow of data across 10-15 years, as with my Unit's work.

The parallel finite corpora which are currently available sit thirty years apart, a span which has been argued (Mair 1997) to be appropriate to reveal significant shifts in language use. Under these conditions, Mair and others (see. Hundt 2006; Leech and Smith 2006) have, in addition to other insights gained, for example into American and British English variation, been able to investigate and substantiate their theories of language change in relation to grammaticalisation, importantly linking this with increased colloquialisation across the period. A limitation of the 'gapped' approach, as pointed out by Mair, is the improbability of capturing actual moments of change (at least, conclusively). Conversely, the

unbroken diachronic corpus of news text has the disadvantage that it covers a time-span too brief to reveal much if anything of grammatical change, but it does reveal lexical and morphological innovation and fashion, and hints of change at the lexico-grammatical levels. It is more likely to capture many actual moments of journalistic invention (though again unverifiably), as well as the early quotation and exploitation of neologisms.

The notions of studying change and of processing text chronologically, tracing neologistic activity, were and probably still are considered significant philosophical and methodological steps forward. However, by the 1990s, the computational capacity to store and retain all past data led to a new theoretical breakthrough and associated methodologies. It allowed the bi-directional processing of text[8], and thus the analysis of character strings as multiply segmentable. This breakthrough was explored in the APRIL project, 1994-1997. The project concerned the identification and classification of hapax legomena and other neologisms at point of first-occurrence (all singletons at that stage) in the journalistic text flow, with the purpose of understanding the nature of the lexicon at its most productive level. It became clear that while a new word sometimes consisted of existing words or affixes in new combinations, the parse for new formations yet to appear in dictionaries very often revealed ambiguity. Multiple analyses of the character string for two or more possible points of segmentation were necessary. Words could not simply be parsed from left to right, but required crab-wise incremental assessment, whereby the word was regarded as potentially constituted of a series of partially overlaid sub-strings, and these analytical possibilities were ranked by 'cost' or likelihood. Such recursion was now achievable with the application of a novel 'chart parser', modified to operate at character level.

A host of terminological issues accompanied the move into diachronic corpus study. An early question concerned what the terms *corpus*, *database* and *text archive*. By 1990, a *corpus* was still a designed, finite collection of partial or whole raw texts (if annotated, then an *annotated corpus*), and processed electronically, typically as a synchronic entity. A *database* could encompass a *corpus*, but typically denoted a collection of knowledge or facts, the products of raw corpus analysis; and a *text archive* was a catalogued text collection, whether or not prepared for processing as a corpus. Inevitably, there remains some overlap in the use of these terms.

Another terminological distinction exists between *historical* and *diachronic* (Renouf 2002). Historical linguists use the term *diachronic* to describe their study in a collective sense, since as a body of scholars, they study the whole realm of text across time, even though in principle each individual investigation could be focussed synchronically in the past. To modern diachronic linguists, *diachronic* is used in contrast to *synchronic*; while *diachronic* can also mean 'across time' to both historical and modern corpus linguists.

Moreover, when historical linguists speak of *diachrony*, their mental time-frame is likely to be calibrated in centuries and millennia. Modern diachronic linguists, in contrast, are typically comparing text across time-frames of ten to thirty years. Rissanen (2000) has referred to the longer time-frame as 'long diachrony'; Kytö, Rudanko and Smitterberg (2000) have talked of the shorter time-frame in terms of 'short-term change in diachrony'; while Mair (1997) has dubbed this shorter time-span 'brachychrony'.

The gradual coming together and overlapping of the periods of text studied by historical and modern corpus linguists has also given rise to a terminological lacuna for the period extending from 1900 to today. 'Late Modern English' would normally seem an appropriate term, but of course this means something much earlier to the historical linguist. In fact, corpus linguists refer to this nameless period as anything from 'modern' to 'present-day' to 'current' to '20th century' and '21st century'.

4.3 Practical issues concerning the modern diachronic corpus

For some types of modern diachronic corpus study, such as that based on the comparison of two parallel designed ('standard') corpora like *Brown* and *Frown*, the parameters of corpus text selection can be set according to specific linguistic criteria. For diachronic study within an unbroken flow of text, the nature of the corpus is dictated by the electronic data which are available; in the AVIATOR, ACRONYM and APRIL projects, this was quite simply national news text, although this did not prevent us, for example, from branching out to compare English and French news streams. The scope of the diachronic corpus is also dictated by the time-span of available data. In 1990, we in my Unit began the AVIATOR project with a store of just 5-6 years of text; now in 2004, we have over 15 years' worth of *Independent* and *Guardian* news text. Another problem is that the phenomena that are typically under scrutiny in modern diachronic study, such as first occurrences, neologistic usages and new patterns, dribble through in ones and twos when analysed diachronically. Very delicate statistical measures may be introduced gradually, but their performance on sparse early data has to be scrutinised closely.

4.4 Was building the modern diachronic corpus worth the effort?

Yes, it was. The monitoring of text reveals innovation, trends, vogues, changes, revivals, patterns of productivity and creativity. In the ACRONYM project, for instance, we have been able to discover changes in sense relations in text over time by monitoring the change of collocational profiles of two words. For the word *cleansing*, we find an upsurge of activity accompanied by a change in sense relations (and thus incidentally reference) in the latter part of 1992 in *Independent* news text. This is shown in Figure 5.

ind8901-ind9203 inquiry

centre	massacres	slaughter
	genocide	war
	atrocities	offensive
	killings	refugees
	expulsions	shelling
ind9204-ind9912	repression	bombardment

Figure 5: Ranked list of 'nyms' for the term *cleansing* in *Independent* news text, July-December 1992

Meanwhile, in the APRIL project, the analysis of diachronic data allows us for the first time to trace over time the neologisms formed with particular productive affixes. Figure 6 illustrates graphically a significant growth and subsequent decline in the frequency of neologisms formed with the prefix *e-*.

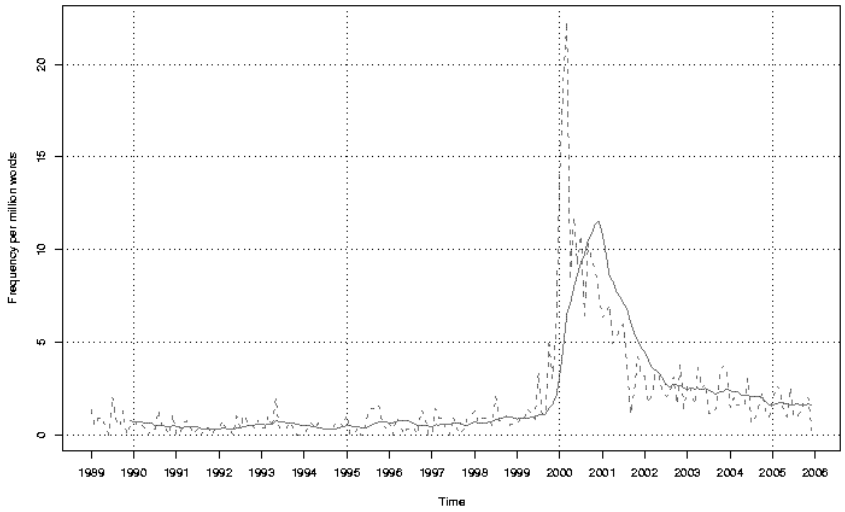

Figure 6: Growth and decline in number of new words with prefix '*e-*' in *Independent*/*Guardian* news text, 1989-2005. Dotted line is frequency per million words; solid line is a moving average

In Figure 7, we illustrate this with some output for the prefix *cyber-*. The data are the result of morphological analysis by a specially-modified 'chart parser'. By focussing at sub-word level in our diachronic text, we have a means of grouping and classifying these hapax legomena of text. In Figure 8, we demonstrate how it is then possible to extrapolate from the same diachronic corpus of news text an indication of the ranking and relative growth (for the prefixes *techno-* and *cyber-)* and drop (for the prefix *poly-*) in productivity for affixes over a period of ten

years. These kinds of cumulative analysis provide us, in principle, with a basis for the prediction of the overall structure of the lexicon.

Query Results from *The Independent* - All of 1999

New words with the prefix **'cyber'**

word	parse	tag	month
- cyberstalker	cyber- (stalker)	NN1	9901
- cyber-future	cyber- '-' (future)	NN1	9901
- cybermemoir	cyber- (memoir)	NN1	9901
- Cybersouls	cyber- (souls)	NN2	9901
- Cybertalk	cyber- (talk)	NP1	9901
- cyber-event	cyber- '-' (event)	NN1	9901
- cybergeek	cyber- (geek)	NN1	9901
- cybermemoirists	cyber- (memoirists)	VVZ	9901
- cybersong	cyber- (song)	VVG	9901
- cyber-sessions	cyber- '-' (sessions)	NNT2	9902
- cyber-glow	cyber- '-' (glow)	NN1	9902
- cyber-realist	cyber- '-' (realist)	NN1	9902
- Cyber-Valentine	cyber- '-' (valentine)	NP1	9902
- cyber-commerce	cyber- '-' (commerce)	NN1	9902
- cyber-auction	cyber- '-' (auction)	NN1	9902

Figure 7: Neologisms with the prefix '*cyber-*' in *Independent* news text, 1999

	89	90	91	92	93	94	95	96	97	98	99
non	1	1	1	1	1	1	1	1	1	1	1
un	1	1	1	1	1	1	1	1	1	1	1
anti	1	1	1	1	1	1	1	1	1	1	2
ex	1	1	1	1	1	1	1	1	1	1	2
re	1	1	1	1	1	1	1	1	1	1	2
pre	1	1	1	1	1	2	1	2	2	2	2
post	1	2	2	2	1	1	1	2	2	1	2
over	1	1	1	1	2	2	2	2	2	2	2
euro	1	2	2	2	2	2	2	2	2	2	2
super	2	2	2	2	2	2	1	2	2	2	2
mini	2	2	2	2	2	2	2	2	2	2	2
pro	2	2	2	2	2	3	2	2	3	3	2
semi	2	2	2	2	2	3	3	3	3	3	3
out	2	2	3	3	3	3	3	3	3	3	3
sub	2	2	3	3	3	3	3	3	3	3	3
under	2	2	3	3	3	3	3	3	3	3	4
multi	3	3	3	3	3	3	3	3	3	3	4
mid	2	3	3	3	4	4	3	3	3	4	4
pseudo	3	3	3	4	4	4	3	3	4	4	4
quasi	3	3	3	4	3	3	4	4	4	5	4

	89	90	91	92	93	94	95	96	97	98	99
mega	4	4	4	4	3	3	3	4	4	4	4
inter	2	3	3	4	4	4	4	4	4	5	5
micro	3	4	4	4	4	5	4	3	4	4	4
ultra	4	4	4	4	3	4	4	4	4	4	4
counter	3	3	4	4	4	5	5	5	5	5	5
eco	4	5	4	4	5	5	5	4	4	4	5
hyper	3	5	5	5	5	5	4	4	4	4	5
mis	2	3	4	4	5	5	5	5	5	5	6
neo	3	4	4	4	5	5	5	5	5	4	5
auto	4	4	5	5	5	4	5	5	4	5	5
bio	4	5	4	5	5	5	5	5	5	4	5
mock	4	4	5	5	5	5	5	4	5	5	5
dis	3	4	5	4	5	5	5	5	6	6	6
techno	6	7	6	6	5	5	3	3	4	4	5
tele	5	5	5	5	5	5	5	5	5	5	6
cyber	8	9	8	9	8	5	3	2	2	2	3
arch	5	5	6	6	5	7	5	6	6	5	5
trans	4	6	5	6	5	5	6	5	5	7	7
proto	5	5	6	7	6	7	5	5	6	5	5
poly	4	5	5	5	6	7	6	6	6	6	7

Figure 8: Prefix banding of productivity for the prefixes *techno-*, *cyber-* and *poly-* in *Independent* news text, 1989-1999.

5. The evolution of the cyber-corpus

I use the term 'cyber-corpus' to refer specifically to texts on the World Wide Web treated as an on-line corpus, in the sense of functioning as a source of language use, of instances of language use extracted from the Web and processed to provide data similar to the concordanced and other analysed output from a conventional corpus.

5.1 Drivers for the Web as corpus

There are three drivers for the treating the texts on the Web as a source of linguistic information. The primary one is probably *serendipity*. The Web emerged in the 1990s, and though it was created to store textual and other information, people gradually realised that it contained useful and otherwise unavailable linguistic data, new and rare words not found in existing corpora, as well as more varied types of language use. *Pragmatic* considerations were the secondary driver. Corpora are very expensive and time-consuming to build, so that they are limited in size, out of date by the time of completion, and they do not change or keep up with language change. Web texts, on the other hand, are freely available, vast in number and volume, constantly updated and full of the latest language use. *Theoretical* curiosity about the nature and status of rare, new, and possibly obsolete language phenomena had been nurtured by many a linguist prior to the birth of the Web, and as web texts began to accumulate, corpus linguists began to negotiate the commercial search engines in an attempt to search for instances and counter-instances to test their new or long-nurtured hypotheses. So the demand was there, and as with the opening of any new highway, demand increased simply in response to its existence.

5.2 Theoretical issues concerning the Web as corpus

The theoretical objections to using the Web as a corpus come thick and fast. The problem is exacerbated where the processing, as with *WebCorp*, is carried out in real time. One of the many issues is the uncontrollability of the data, which form an arbitrary and instantial corpus that changes like the sand with each new search. Another is that fundamental corpus-linguistic methods such as exhaustive and quantitative study are impossible or inhibited in this non-finite context, where the total population is not known or knowable, and the significance and interpretability of results are thrown into question. A fuller range of problems has been explored, e.g. in Kehoe & Renouf (2002), Renouf (2003), Renouf et al (2006), and is echoed in the practical issues enumerated below.

5.3 Practical issues concerning the Web as corpus

Many linguists have not yet reconciled themselves to the advantages of accessing the Web as a corpus, finding all manner of objections to it in principle. Chief among their concerns are probably the heterogeneity and arbitrariness of the text and hence the status of the language use found on the Web. These concerns are shared by us as active web-as-corpus providers, who struggle to extract from the Web and process them to a degree which at least approaches the interpretability and usability of the output from conventional diachronic corpora, but which currently fails to match the quantifiability of conventional finite corpora. Researchers like ourselves cope with such issues as the state of Web text, with its typographical errors and erratic or absent punctuation (useful for sentence identification); the heterogeneity of web-held data, the handling of web pages with their hotchpotch of more and less text-like texts; the need for speed of search, retrieval and processing; language identification; and the absence of a established standard for dating texts, with particular reference to the date of authorship (whence also the impossibility of achieving reliable chronological text sequencing for diachronic study).

5.4 Was building the Web as corpus worth the effort?

Yes, it was. *WebCorp* gives us the possibility of retrieving instances of words and phrases in text that are either too rare or too recent to appear in conventional text corpora. Figure 9 illustrates the case with reference to the term *Enron*, which appeared on the Web as soon as the scandal broke, and almost immediately became productive, spawning not just the search term *Enronomics*, but also *Enronyms*, *Enronitis*, *Enronity*, *Enronethics*, *Enronizing*, *enronish*, *Enronitize and enronomy*, to name just those variants in our sample.

WebCorp output for search term "Enronomics"
Domain: ".uk OR .com"

1. attack Bush's economic policies with the term "Enronomics" (a phrase that apparently originated in a
2. to Believe He Knows About the Economy? Enronomics = Contributors Get Richer 1/16 Message to
3. corporate malfeasance. Recently spotted Enronyms: Enronitis, Enronify. Enronomics. silver bullet: In war, it's an
4. is laid bare by what rivals call 'Enronomics' - the political fable of the Enron corporation
5. Dems slogan for slogan and neutralize the Enronomics accusations, may I coin the term "Enronethics
6. C.) 2 p.m. Breakout Workshops -- Confronting Enronomics -- Arianna Huffington, Rep. George Miller (D-Calif.)
7. investigators Wed. (DBN Subscription Required) Democrats knock 'Enronomics' - But strategists warn against
8. a political problem for TeamBush-with talk of "Enronomics," or "Enronizing" Social Security and Medicare. But
9. believing their press, watch out. It's Enronomics, folks. The rich seducing the poor, while
10. Riseth The Conservative Cliterati What Monica Cost Enronomics Catholics and Condoms Virtual Rape Ring-a-Ding
11. national energy policy based on the same Enronomics as its own disastrous business strategy. But
12. people, to be enronish and to practice Enronomics. "We've seen ugly, enronish sights before," Jane
13. The Looting of America: Reaganomics, Clintonomics and Enronomics AL MARTIN is America's foremost
14. strategy") http://www.dailyhowler.com/h012902_1.shtml Enronomics Explained (deliberately driving the country into
15. who's spent two weeks talking about Bush's "Enronomics" and "Enronizing" Social Security. He capitulated to
16. McFedries said. He placed worse odds on "Enronomics," reminiscent of "Reaganomics," sticking. "The Democrats
17. ideology. It blows the lid off Bush's Enronomics, and his plan to Enronitize Social Security
18. at alarming rate. ENRON (I call it Enronomics) phenomena - soft money manipulating the policies and
19. can only be described as trickle down enronomics! That's exactly what Bush is doing to
20. hardest hit by the Bush trickle down enronomics. Now it looks like the Bush enronomy

Figure 9: *WebCorp* output showing the productivity of *Enronomics,* May 2004.

WebCorp output for search term "medicalisation"

1. legislation was shifted from criminalisation to medicalisation of drug use. Public demand for effective treatment
2. Sheldon examines the causes and effects of the medicalisation of abortion, focusing on the role that law
3. to the lawfulness of the procedure remains. The Medicalisation of Abortion and the Common Law The
4. celebrated 6. 02 6620 2970 Fax: 02 6620 2161 The Medicalisation of Sexual Violence: The Social and Political
5. their frustration and disappointment with the increasing medicalisation and intervention in maternity care and
6. began which medicalise, and therefore pathologise, difference. The medicalisation of epilepsy has inevitable
7. 28. Grubb, A.; " Abortion Law in England: The Medicalisation of a Crime" , 18 Law, Medicine and Health
8. School: School individuals and society; autonomy and paternalism; the `medicalisation' of life; the goals of
9. medicine; the society is: adapt yourself!" (Touraine) The psychologisation / medicalisation of school education
10. is a strategy for to turn back the tide on the 'medicalisation' of everyday life. People, who would previously
11. upon initially vague ideas of clinical stress, medicalisation of tension, emotional fever, & on perceptions about
12. and death and to exert control, in line with most medicalisation of childbirth, but Petchesky points out that
13. violation and we refuse to see the medicalisation of something that is wholly unnecessary. We Sociology,
14. Brandeis University, Massachusetts Medical sociology The medicalisation of deviance Modern genetics Email:
15. Crawford, E on midwifery knowledge before the NHS, the medicalisation of childbirth and on the teaching of
16. not to be underestimated.[34] A danger of medicalisation of the law becomes apparent, for instance, in
17. wards etc Centre, Oxford University. Leading critique of the medicalisation of distress via the diagnosis of
18. PTSD the postcolonial condition citizenship and rights discourse medicalisation of the legal subject revisiting
19. consent AIDS A political sociology of lifestlye pharmaceuticals and medicalisation. University of Sussex Mr
20. MM Hopkins THE Midwifery. Professional education Fiona Dykes Infant feeding. Medicalisation of childbirth
21. Norma Fryer Ethical issues relating of mental illness. Experts call it the "medicalisation of human distress" –
22. the trend to treat professionalisation' of lay researchers and representatives. The Medicalisation of Health and
23. Use of lethal injection and the general medicalisation of killing are in direct conflict with Field D (1994)
24. Palliative medicine and the medicalisation of death, European Journal of Cancer Care Recentering Class and C

Figure 10: *WebCorp* output showing old and new uses of *medicalisation,* 2004.

Figure 10 also exemplifies the richness of web text in terms of the changing use of words across time which it (inadvertently) yields. Here, the term *medicalisation*, in the new sense of 'treating as a medical condition the natural facets of life', is found alongside the established meanings of 'decriminalising (of drug use)' and 'treating terminal illness by palliative means'. In addition, this text shows how a derived term like *medicalisation* can spawn parallelisms, here *pathologise*, *psychologisation* and *professionalisation* (or vice versa). Searching the Web with the aid of wildcards and brackets, combined to represent complex patterns, yields another dimension of valuable information. Figure 11 presents an extract of the pattern e.g. *dr[i|o]ve[s|n|] * [a|]round the*.

1. Start up *drives me round the* twist
2. Fury over lorry that *drives residents round the* bend
3. Over used, that stupid drumbeat *drove me round the* bend
4. We quit – you've *driven us round the* bend
5. The noise *drove her around the* bend
6. Her Majesty *was driven twice round the* Mews yard
7. 'Sick' Diana pic *drives critics round the* Benz

Figure 11: *WebCorp* results for pattern *dr[i|o]ve[s|n|] * [a|]round the*, using wildcard option and bracketing.

6. A possible future of the corpus in corpus linguistics

While the range of small, medium and large corpora will undoubtedly continue to grow, it seems likely that the World Wide Web will also figure largely in future corpus development, both as a source of data (whether online or downloaded) in itself, and also as a repository for datasets and text collections, a staging post in the location and access of externally-held intranet text archives and corpora, and a medium in which researchers can cooperatively create, process and share corpora. This prediction is based on discussion that has been ongoing for several years now concerning the post-Internet era of 'GRID'-like technology.[9] The GRID[10] refers to a set of new hardware that will support, in principle cooperatively, a more distributed processing environment.[11] To the corpus linguist, the new web infrastructure, as it gradually appears, may not seem much different from now.[12] The new structure will probably be a layer of protocols or routines sitting on top of existing web architecture which function in a similar way to already familiar protocols such as 'ssh' (secure shell, for using a computer remotely); 'email' (Pine, Mulberry etc, for sending electronic messages); and 'ftp' (file transfer protocol, for copying files over networks).

6.1 Drivers for the corpus of the future

As has been implied, the driver for this development towards co-operatively processed corpora will be a *pragmatic* one. As the scale and type of computing

increases, there will be a growing need for regionally distributed computing, for information and resource sharing, and of handling the logistics of moving around distributed data and other resources.

7. Concluding remarks

I have in this paper traced the development of corpora 'twenty-five years on', reviewing the issues involved at each stage. I have shown that as the years have passed, the design of corpora has continued to be characterised by the tension between the desire for knowledge and the constraints of practical necessity and technological feasibility. The vision of corpus evolution that I have presented, being a chronological overview, has tended to coincide with major milestones in technological evolution. This is inevitably a simplified picture, which has not given due attention to other factors. Corpus linguists themselves, while still curious about real language in use, have been growing increasingly aware of what it is, and aware that every type of corpus is in principle now available to them, or possible to construct.

As computing technology hurtles on, it does open up access to new corpus resources and the possibility of new orders of corpus magnitude such as I have inventoried, but it also provides the infrastructure to support smaller-scale, more intricate corpus design and organisation, exemplified by the multiply-layered, cross-disciplinary databases such as the *Corpus of Early English Correspondence* (Nevalainen 2003), the *Corpus of Early English Medical Writing* (Taavitsainen 1997) and the *LAOS* (*Linguistic Atlas of Older Scots*). And much research activity has not evolved directly in response to the availability of large-scale computing resources but to the linguistic needs of the corpus linguist. The consolidation of existing corpora is, for example, the focus of much attention. Synchronic corpora such as the *BNC* are being renewed or updated; diachronic and variational collections such as the *Archer Corpus* are growing incrementally*;* corpus sets like the *ICLES* learner corpora and the *LOB, FLOB, Brown, Frown* stable are being extended.

Ultimately, each of these corpus-oriented activities, whether ground-breaking or incremental, computational or linguistic, theoretical or intuitive, progressively enriches our understanding, whilst broadening the scope of enquiry for the next phase.

Notes

1 Thus, when I speak of the early 'small corpora', I am referring primarily to the pioneering work of the first generation of corpus linguists in the form of the *Brown Corpus* (1961-1967), *Survey of English Usage* (1953-1987) *Corpus* (1964-), *Lancaster-Oslo-Bergen Corpus* (1960-1967) and *London-*

Lund Corpus (1970-1978); and not so much to the important but second-wave small corpora of specialised text, which were created from 1980 onwards, notably those of historical English such as the *Helsinki* and *Archer* corpora; the corpora of regional varieties of English, including Somerset/Norfolk and Manx English dialects; the corpora of international varieties of English such as those of the *ICE* Project; the corpora of Learner English such as those of the *ICLE* and *CLEC* projects; the multilingual corpora such as those developed in Sweden and Norway; or the specialised technical corpora, such as *MICASE Corpus*.

2 The *BNC* was created in 1991-1994 by an academic-industrial consortium whose original members were: Oxford University Press, Longman Group Ltd, Chambers Harrap, Oxford University Computing Services, the Unit for Computer Research on the English Language (Lancaster University), British Library Research and Development Department.

3 Authenticity was considered essential, though initially what precisely constituted authentic text, and what its benefits were to language description, were not initially questioned in detail. Sinclair remedied this, in his article on 'naturalness' in language, published along with a reply by Owen, in a useful volume on the topic (1988). Meanwhile, whilst it was clear to most corpus creators from the outset that dramatic dialogue did not constitute an authentic use of speech, and was thus to be excluded, the issue of how to classify conversation within novels has remained largely unresolved.

4 AVIATOR stands for 'The Analysis of Verbal Interaction and Text Retrieval'.

5 ACRONYM stands for 'The Automated Collocational Retrieval of Nyms'.

6 APRIL stands for 'The Analysis and Prediction of Innovation in Text'.

7 The semantically unrelated word *punters* creeps into the nymic output by virtue of the unexpected number of collocates it shares with *solace*. It is not possible by statistical means to avoid such oddities entirely, only to find the best variables for minimising their occurrence.

8 The use of short word histories, in the form of tri-grams where two recognised words to its left were used to recognise a third unknown node word, preceded this work (Jelinek 1976), but this was in the era of off-line processing of static finite corpora, rather than the new 'on-the-fly' processing of our approach.

9 Particle Physicists devised the World Wide Web, and facing many of the problems the 'Grid' aims to solve, took the first initiative.

10 Other projects which resemble the Grid, such as Internet II and Internet III, exist internationally.

11 The term GRID is "chosen by analogy with the electric power grid, which provides pervasive access to power", a facility which the user simply plugs into, without knowing where the electricity is processed or comes from, and which "has had a dramatic impact on human capabilities and society" (Foster and Kesselman 1999: xix). If and when the Internet saturates, it will need to be re-organised. The result will probably be that the Web contains more indexes to texts than texts themselves, which will be stored in 'DataGrid intranets'. Specialised software will be needed to process them within a predicted system of global distribution of machinery (from regional computing hubs down to domestic machines), and 'Grid middleware' will be required to find and organise them, and to protect security. This new system will also require better text content annotation, of the kind being devised by the 'Semantic Web' community of computational linguists.

12 The definition of large-scale processing to the particle physicist envisages something on a vastly larger scale than that of even the most ambitious corpus linguist.

References

Burnard, L. (ed.) (2000), *Reference guide for the British National Corpus (world edition)* (*BNC* CD-ROM). Oxford: Humanities Computing Unit of Oxford University.

Foster, I. and C. Kesselman (eds.) (1999), *The Grid: blueprint for a new computing infrastructure*. San Francisco: Morgan-Kaufmann.

Fries, U. (1993), 'ZEN-Zurich English newspaper corpus', in: Kytö, M., M. Rissanen and S. Wright (eds.) *Corpora across the centuries. Proceedings of the first international colloquium on English diachronic corpora*. St Catharine's College, Cambridge. Amsterdam/Atlanta: Rodopi. 17-18.

Hundt, M. (2006), '"Curtains like these are selling right in the city of Chicago for $1.50": the mediopassive in 20th-century advertising language', in: Renouf, A. and A. Kehoe (eds.) *The changing face of corpus linguistics: papers from the 24th international ICAME conference*, Guernsey, 23-27 April 2003. Amsterdam/New York: Rodopi. 163-184.

Izre'el, S. and G. Rahav (2004), 'The Corpus of Spoken Israeli Hebrew (CoSIH); phase I: the pilot study', in: Oostdijk, N., G. Kristoffersen and G. Sampson (eds.) *LREC 2004: fourth international conference on language resources and evaluation; workshop proceedings: compiling and processing spoken language corpora*, Lisbon (Portugal). 1-7.

Jelinek, F. (1976), 'Continuous speech recognition by statistical methods', *Proceedings of the IEEE*, 64 (4): 532-556.

Johansson, S. (ed.) (1982), *Computer corpora in English language research*. Bergen: NAVF.

Kehoe, A. and A. Renouf (2002), 'WebCorp: applying the web to linguistics and linguistics to the web', in: Lassner, D., D. DeRoure and A. Iyengar (eds.) *WWW2002 conference*, Honolulu (Hawaii). New York: ACM. Available from: http://www2002.org/CDROM/poster/67/

Kurzweil, R. (1990), *The age of intelligent machines*. Cambridge (Massachusetts): M.I.T. Press.

Kytö, M., J. Rudanko and E. Smitterberg (2000), 'Building a bridge between the present and the past: a corpus of 19th-century English', in *ICAME journal*, 24: 85-97.

Leech, G. and N. Smith (2006), 'Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English', in: Renouf, A. and A. Kehoe (eds.) *The changing face of corpus linguistics: papers from the 24th international ICAME conference*, Guernsey, 23-27th April 2003. Amsterdam/New York: Rodopi. 185-204.

Mair, C. (1997), 'Parallel corpora: A real-time approach to the study of language change in progress', in: Ljung, M. (ed.) *Corpus-based studies in English*. Amsterdam/Atlanta: Rodopi. 195-209.

McCarthy, M (ed) (1988). *Naturalness in Language*, ELR Journal (New Series) Vol. 2.

Nevalainen, T. (2003), *Historical sociolinguistics*. London: Longman.

Nevalainen, T. and H. Raumolin-Brunberg (1996), 'The corpus of Early English correspondence', in: Nevalainen, T. and H. Raumolin-Brunberg (eds.) *Sociolinguistics and language history. Studies based on the corpus of Early English correspondence*. Amsterdam/Atlanta: Rodopi. 38-54.

Owen C. (1988): 'Naturalness and the Language Learner', in McCarthy 1988, 21-46.

Renouf, A. (1987), 'Corpus development', in: J. McH. Sinclair (ed.), 1-15.

Renouf, A. (2002), 'The time dimension in modern corpus linguistics', in: Kettemann, B. and G. Marko (eds.) *Teaching and learning by doing corpus analysis. Papers from the 4th international conference on teaching and learning corpora*, Graz, 19-24 July 2000. Amsterdam/Atlanta: Rodopi. 27-41.

Renouf, A. (2003), 'WebCorp: providing a renewable data source for corpus linguists', in: Granger, S. and S. Petch-Tyson (eds.) *Extending the scope of corpus-based research: new applications, new challenges*. Amsterdam/ Atlanta: Rodopi. 39-58.

Renouf, A., A. Kehoe and J. Banerjee (2006), 'WebCorp: an integrated system for web text search', in: Nesselhauf, N., M. Hundt and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi. 47-68.

Rissanen, M. (1992), 'The diachronic corpus as a window to the history of English', in: J. Svartvik (ed.) *Directions in corpus linguistics: proceedings of Nobel Symposium 82*, Stockholm 4-8 August 1991. Berlin: Mouton de Gruyter. 185-205.

Rissanen, M. (2000), 'The world of English historical corpora', in *Journal of English Linguistics*, 28 (1): 7-20.

Salinger, J. D. (1951), *Catcher in the Rye*. Boston: Little Brown & Co.
Sinclair, J. McH. (ed.) (1987, *Looking up*. London/Glasgow: Collins ELT.
Sinclair, J. McH. (1988), 'Naturalness in language', in McCarthy 1988, 11-20.
Summers, D. (1993), 'Longman/Lancaster English language corpus – criteria and design', *International Journal of Lexicography*, 6 (3): 181-208.
Taavitsainen, I. and P. Pahta (1997), 'The corpus of Early English medical writing', *ICAME Journal*, 21: 71-78.