The Time Dimension in Modern English Corpus Linguistics

Antoinette Renouf

Abstract

The corpus-based analysis of modern English tends to focus on language which has been written or spoken at a particular point in time, and a corpus is *conventionally set up as synchronic entity. A synchronic study* is *often entirely appropriate, but language* is *a changing phenomenon, and linguists are also interested in that dimension: curious to trace an earlier language feature through to the present, or a current feature back to its source, and in studying recent changes in language use.*

Within this context, I shall discuss new developments in three areas of research activity: firstly, the setting up of a means of tracing morphological, lexical and semantic changes in Modern English text across time; secondly, the use of the web as a linguistic resource; and thirdly, the coordination of methodologies and resources in modern and historical corpus linguistics.

1. Introduction

At the 'Early Dictionary Databases' conference in Toronto in 1993, I reported on the A VIA TOR project, which was reaching completion within my unit, then at Birmingham (Renouf 1994). The purpose of AVIATOR was to develop an automated system to identify and record ongoing lexical change in modem English text. I began:

> The era of the computerised corpus has arrived. Computing technology has developed rapidly, allowing collections of source data to be held and accessed electronically. Such a data store can be very large indeed, and added to easily [...] With the growth in computer storage capacity has come text processing software, capable of carrying out exhaustive searches at very high
> speeds [...]

Whilst great strides had indeed been made in the new field of modem English corpus linguistics by 1993, a far more sophisticated state of affairs obtains today. There is now virtually no technological limit to what can be done in the way of creating and exploiting textual corpora, and things are moving fast. In the last decade, alongside the study of 'general' modem English through corpora such as the BNC, all manner of studies of variation - notably of region (e.g. Greenbaum/ Nelson 1996) and learner language (e.g. Granger 1998) - have been set up.

But one area has not moved. Back in 1993, I continued:

> This technology has been developed to assist in the production of synchronic accounts of the language, and the source data is typically treated as a static entity, a window at a given point in time [...] A simple modification would be to order the citations according to first and subsequent occurrence, which would allow a diachronic study within a bounded, finite corpus [...].

Nevertheless, with the exception of the work of my unit and that of Mair's team in Freiburg, the focus for corpus provision and study in current-day English, remains fixed on language at a point in time. Yet language is a changing phenomenon, simplifying and becoming more complex in varying response to the changing world. Seminal studies of language change exist. Why then, eight years on, is there still so little specific provision for the diachronic corpus-based study of modem English? A number of reasons suggest themselves.

The obvious reason is that synchronic descriptions of the language are vital not just in themselves but for a whole range of academic and commercial applications which require knowledge of the current conventions of language use in the English-speaking community. They thus have primacy and are best resourced. Furthermore, the fact is that the language is both static, in the sense of being in a particular state, and dynamic; at any moment in history, text is being constructed according to a generally accepted if fuzzy-edged set of conventions, whilst at the margins, new usage is creeping in and obsolescent and ephemeral items are dropping out. The community has quite reasonably chosen to focus on the stability rather than the movement.

But there are other inhibitors - political, psychological and practical - to progress in modem diachrony. As early as 1982, Sinclair referred to the future possibility of "vast, slowly changing stores of text," providing "detailed evidence of language evolution" (1982), yet this vision has been slow to take hold. For some individual linguists, there is still novelty in being able to study real text *per se,* to investigate previously inaccessible areas such as collocation and vocabulary, while for the better resourced, there may nevertheless be an inertial barrier to diachronic corpus study, as there was in the 80s, to working with finite corpora. There is also a delay in cross-disciplinary fertilisation: many linguists take diachrony for granted in relation to Earlier Englishes, but do not seem to make the connection for modem English. There may be a delay in self-definition, exacerbated by terminological barriers: an expert in language change will not necessarily see him/herself as a 'diachronic linguist', and not know that dynamically-processed corpus data would make hislher task so much more rewarding. There may be an unclarity as to what diachronic linguistics entails; some linguists study language change in synchronically-processed corpus data, perhaps subconsciously assigning to personal intuition the role of pre-corpus or post-corpus point of comparison. Computational linguists will extract new terms or monitor lexical acquisition (Boguraev/Pustejovsky 1996, Fairon 2000) from vast amounts of electronic text, yet not employ a fully diachronic methodology.

A change in perception is necessary to stimulate the move to modem diachronic corpus study, but it is not sufficient. The financial resources are not generally available for the ongoing handling of text, which remains beyond the means of individuals and which needs careful cost-benefit justification for industry. New large-scale corpus projects could take the initiative, but large investors tend to tread warily and slowly. So, one way and another, the necessary infrastructure has not yet fully emerged. This is the background to an exposition on modem diachronic corpus linguistics, which will begin with a definition of the terms involved, move on to enumerate some types of linguistic discovery that can be made in text over time, and then present a series of diachronic corpus models, examining recent developments in each as possible ways forward for modem diachronic corpus linguistics (which I shall refer inelegantly to as *MDCL).*

2. Defining the object of study

2.1 Definition of 'language change'

To a corpus linguist, language change is that change which is identifiable and measurable within an existing corpus of text of a particular domain or variety. It concerns the birth, life and death of elements of language, ranging from morpheme to phrasal unit to clause, in text across time. Change manifests itself in new coinage, in the spread of a feature, in patterns and degrees of productivity, in the gradual assimilation of a new feature into the conventional lexicon, or its eventual departure. Changes can operate at any level of textual organisation: lexical, lexico-grammatical and grammatical; semantic, referential, functional, pragmatic, sociolinguistics, and so on.

Mair has observed (2000: 196), quoting Lass (1980:95), that it is impossible to observe the exact moment of change, particularly in speech, but that the inference of change is feasible. To put this in context, Mair is primarily referring to slower types of language change, within grammar or lexico-grammar, where the 'spread' (increased usage) of a feature may be observable, but where it is hardly sensible to talk of the precise moment of change. He is also referring primarily to his own corpus resources, namely small parallel corpora separated by a 30-year gap, where the very first formulation may well have occurred outside or between the corpora. So what can be observed depends on the corpus resources available, the definition of change, and the linguistic feature under scrutiny. In a situation like ours at Liverpool, where the corpus is a dynamic flow of journalistic text, where 'coinage' is defined as the first manifestation of change, and where the language feature is in the faster-evolving area of lexis, it is possible to observe the birth of a new lexeme; what is not possible is to know for sure that this is what one is observing.

2.2 Definition of 'diachronic linguistics'

The study of language across time, diachronic linguistics, is already intrinsic to the corpus-based study of Early Englishes. For historical corpus linguists, this is a term referring to the study of Earlier Englishes as a whole, embracing both

individual studies with a synchronic focus, and those dealing with change through and across time periods. Recently, historical corpus linguists have begun to differentiate between the traditional scope of their field, which covers the centuries and is referred to as *long diachrony* (Rissanen 2000:9), and "the recent scholarly interest in 'short-term change in diachrony"' (KytO/Rudanko/ Smitterberg 2000:85) covering the last century, which they see emerging. For modem corpus linguists, diachronic linguistics is typically the study of change in one or more aspects of language use just within (or across) a timespan of 10-30 years, a relatively brief space of time that Mair (1997) has termed *brachychrony*.

2.3 Definition of 'modern English'

Both historical and modem corpus linguists refer to the object of study variously as *present-day*, *current* and *twentieth-century* English. But there has been no real discussion or consensus as to the particular point or period in history that this occupies. There would be no ambiguity if contemporary corpora could be built and analysed immediately, so that both source text and descriptive perspective were indisputably set in the present day. But like Samuel Johnson, who selected his source texts to reflect the language use of a past golden age, we still create our corpora retrospectively (albeit from necessity rather than choice), and we have not yet dealt with the question of time-frame other than impressionistically.

What constitutes modem or present-day English inevitably shifts with time. The point at which it culminates is ultimately today, but there is no consensus over where it begins. For instance, back in 1980, for the Cobuild project, I decided that there had been a sea change in society, and thus in language, with the advent of the Beatles, pop culture and teenage power, and so defined modem English as generally being anything published after 1960 (1987:2). The Birmingham Collection of 18 million words thus covered the period 1960-1986, while the Bank of English assimilated the Collection and is still growing. Mair (1997:203) came to a similar conclusion for his 90s FLaB and Frown corpora, seeing "the late 1960s and early 70s, with their student rebellions [...] as the watershed" in social awareness and norms, and thus that compilations of text from 1991 and 1992, thirty years after LOB and Brown, would "fortuitously capture" the "linguistic repercussions." It might be argued that the advent of email, and more recently text-messaging and chat-rooms, as new textual mediums, mark yet further turning points, where the language conventions represent new orders of democratisation and colloquialisation of language use, whole new sets of conventions which characterise the primary means of communication of the internet generation, together with software publications and webzines, and are thus set to spill over into everyday written English.

3. The kinds of language change that can be observed in diachronic corpora

Language is a changing phenomenon, but what precisely is it that is changing, and what time frames are involved? In this section, I shall briefly set out some of the areas of change in modem English that are appropriate, useful and, in principle, amenable to study, and which my unit has investigated, as outlined in Section 4, Model 3.

Change area 1. The coinage of new words and lexical items

Mair has said, as I mentioned previously, that it is impossible to observe a point of change in language use. In absolute terms, this is true, since written corpora reflect almost nothing of what is happening in the world of speech and only a little of the totality of what is being written and published. It is also true that with the two small parallel corpora at Freiburg (see Section 4, Model 2), the first instance of a particular change apparent in the later corpus may have occurred in the gap between the two. But in the monitoring of a long, unbroken stretch of corpus text, at Liverpool it is at least possible to pinpoint the day on which an item first appears, which is in turn a reasonable clue to its being a neologism.

The criterion for newness for our large, chronologically-analysed corpora is 'that which has not been recorded previously in the data'. This species of 'new item' takes various forms. Neologisms can be regular formations - grammatical inflexions and lexical derivations. They can be new coinages, usually new compounds or derivations of existing words and very rarely new inventions, which we can go on to monitor over the years to see whether and how they are assimilated into the language. They may not actually be new but simply stably rare words, those that are available in the long-term lexicon and at one point wander into the corpus, and then recur periodically throughout the lifetime of the corpus. Other newly-recorded items may in fact be revivals, instances of longneglected lexis or usage.

Change area 2. The changing fortune of a lexical item

Candidate new words can be identified and recorded at birth in a chronologically-processed corpus, and each subsequent occurrence logged and dated. Thus the path of each word over the period may be traced. Of course, the findings are limited to the extent of data: an item may disappear, but in fact destined to reappear in corpus data that is not yet available for processing.

Change area 3. The structure of the lexicon

The 'structure' of the lexicon in text is well known. In a large corpus, it is

composed, in roughly descending order of frequency of occurrence, first of grammatical words, core lexical words, dominant technical terms, discourse organising words, stylistic fillers; then the rarer items - derivational, inflexional,

semantic and co-referential variants, many of which fulfil the discourse role of second and subsequent mention of more frequent words; and it ends in the sump of hapax legomena. Folk wisdom used to have it that the top frequency band was unchanging. Domain studies have undermined this certainty; it can also usefully be tested across time.

Change area 4. The meaning and use of existing words

Lexical semantics in text is a syntagmatic phenomenon. In order to study new lexical uses, it is useful to take the Firthian (1953) view: that we can associate a word's meaning with its collocates. We can then deem the new use of a word to be occurring wherever an existing word is accompanied by a change in its collocational patterning. By establishing a 'collocational profile' for each word in a large corpus, when the subsequent instance of a word's environment does not match the established profile, it is possible to monitor that change to see if it is consistent and significant, and if so, to record it as a case of bonafide new use. In this context, the term *use* also encompasses 'sense' and 'reference'. The emergence of a new sense for a word is relatively rare; change in reference is more typical of mainstream language evolution.

Change area 5. Sense relations

Over time, circumstances can change, and a word may develop a new sense or reference. Correspondingly, it will change its semantic partners. Let us take the word *cleansing:* this has traditionally been synonymous with *spiritual purification,* but following the wars in former Yugoslavia, it has taken on the sense of 'murder', and is now synonymous with *genocide.* Similarly, the word *President* was at one point in history co-referential with 'Bill Clinton', but has since shifted its referential allegiance to 'Bush Junior'.

Using the basic notion of collocation profile outlined above, it is possible to identify a change in the sense relationships between words. In text, two words which are synonyms (or indeed antonyms, hyponyms, taxonyms or meronyms) have very similar collocational profiles. When a word changes its meaning, its collocational profile will also change, and the new profile will more closely match that of its new synonyms or otherwise sense-related partners.

Change area 6. Lexico-grammar and grammar

Some changes in lexico-grammar can begin to be identified in a large, unbroken stretch of text covering a decade or more. They normally take the form of syntactic simplification or reduction, as in the case of *provide* or *enable.* The findings of Mair and others over a 30-year period indicate that a decade is insufficient to capture changes in grammar. Grammatical change, and ideally also lexico-grammatical change, both require reliably tagged corpora.

Change area 7. The nature of productivity in text

By productivity is meant the tendency for a particular linguistic feature to generate more of the formations or kinds of formation in which it has been found to occur. A typical example would be the prefix *cyber-*, one of a small coterie of affixes that have become fashionable in the course of the 1990s. They are used not only for their precise meaning, but as pragmatic markers of vogueishness. Tracing the frequency patterns of such morphemes and words across time reveals clear trends for more productive items; classification of the types of productivity can be in terms of morphology, grammar, word formation type, etymology, semantics, and so on. Rarer items, or those which exhibit less growth, can be observed if subsumed into a more general class. For example, productivity for a rare prefix may be almost indiscernible, but grouped with (e.g. semantically) similar prefixes, something can be said about the group overall. Equally, a very common affix will not be expected to manifest any particular growth in productivity, but the degree of the stability of the top frequency band of affixes to which it belongs may be monitored.

4. Setting up a research infrastructure for MDCL

In the context of the types of language change that I have just identified as some of the appropriate concerns of MDCL, I shall now present five generic design models that would create or at least contribute to creating the kind of corpus environment required to support their study.

Model 1. Treating existing corpora as chronological entities

Every corpus is created to answer a particular research question. If the question is about language evolution, the corpus text must cover a sufficient span of time to evidence elements of change. The oeuvre of a prolific author, for instance, would form a corpus suitable for the chronological tracing of his/her artistic development. Some corpora which have been designed to reflect 'modem usage' in English, particularly the new, very large and even open-ended textual databases of English which are being accumulated, do in fact cover a significant time-span. Surprisingly, neither the BNC nor the Bank of English, to name two major synchronic corpora, each covering what must be at least a 10-year spread of text, have been designed as diachronic entities or set up for chronological processing. But corpora such as these could in principle be treated as diachronic resources.

Model 2. Using parallel, static, sampled modern English corpora

Among modem English corpus linguists, an interest in studying change has been growing despite the lack of empirical means properly in place to investigate it. Adhoc measures have been adopted. Back in 1994, Holmes (1994:27) compared the 1986 WNZC New Zealand corpus with the LOB *corpus,faute de mieux,* and

in spite of acknowledging the obvious constraints imposed by the sociolinguistic dissimilarities, felt moved to observe that "the prospect of using corpus data to infer language change over time is an exciting one. It is clearly possible to make suggestive and interesting comparisons between the frequencies of items in corpora of similar size and composition which have been constructed at different points in time".

Holmes pointed to Mair's compilation of the FLOB and Frown corpora at that juncture as a desirable step. And indeed this is one of the two major moves in modem corpus linguistics towards the creation of a principled, tailored infrastructure to support diachronic study (Mair 1997). Mair's Freiburg initiative has created parallel sampled corpora at thirty years' remove from two earlier models. The problem is to establish what it means to monitor change. Mair has set out to 'infer change' through the creation of two small parallel corpora, the 1990s Frown and FLOB collections. These mirror the earlier Brown and LOB corpora, with a 30-year time gap between the two sets. Such corpora offer a good chance of identifying patterns of consistency across the time gap, as well as spreads in usage, and instances of slower types of linguistic change, such as occur more in grammar than in lexis. Mair expects them to make it possible specifically to: test hypotheses about linguistic change; detect changes overlooked in the literature through lexical frequencies, especially of closed-class items; and, in a
different vein, to tackle systematically the "major methodological issue of interdependence between synchronic regional/stylistic variation and genuine diachronic innovation."

Obviously there are limits to the capacity of any corpus to support linguistic study. As Mair (1997:197) says, "it goes without saying that written corpora are useless for the study of sound change: corpora the size of FLOB and Frown are also too small to systematically investigate neologisms and most word-formation processes." But they have obviously been sufficient to allow a certain perspective on the language, since Mair has already made the important discovery that "most changes observed could be interpreted as a result of the colloquialisation of the norms of written English [H'] over the past thirty years." Hundt (1997:135-152) and other users of the Brown, LOB and Freiburg corpora have also demonstrated some of the many benefits of this approach to diachronic study; inventories of the recent research can be found in Mair (1997 :208-9) and (1999: 139-158).

Model 3. Using large, dynamic English corpora

A third approach to MDCL is the one adopted by my unit over the last twelve years, namely to create a single source, large, dynamic collection of English text and to study the language changes which occur within it across time. What can be observed in any corpus is of course inevitably conditioned by the *source data* and its *timespan*, the *methodology* involved, and the *analytical tools* at the linguist's disposal.

- **Size and timespan of dynamic corpora**

In our case, the data currently consists of 400 million words of UK, broadsheet journalistic printed text. The timespan is co-extensive with the availability in electronic form of the Independent newspaper which at the time of writing spans the 11-year period from 1988 to 1999. We have discovered this time span to be too limited for many purposes. It seems that a longer, as yet unquantified stretch is required to identify statistically significant changes in areas other than lexis with any real confidence. As the years progress and the news text continues to flow, however, the informative value of this corpus will inevitably increase.

Mair has said, on the basis of his smaller corpora, that "most changes observed could be interpreted as [...] the linguistic correlate of a general social trend towards greater informality." Similarly, we have found that most language changes in news text of the last 11 years reflect the events, climate and attitudes of that period (though not all of them will be evident in broadsheet newsdata), as viewed from a British perspective. There is much to be studied in relation to sociolinguistic change.

Over the last decade of the 20th century, Britain has experienced major political, social and environmental changes. Milestones have included the advent of New Labour, devolution, teaching assessment, flexible learning, the Lottery, the 'downsizing' of the labour force, the strong pound, steps in EU negotiation, BSE and other health crises, privatization of the utilities. In common with the rest of the world, Britain has experienced the breakup and concomitant conflict and brutality of several countries, global warming, the growth of the Internet, and the prospects for the new millennium.

Such events awaken, particularly in journalists, social responses of enthusiasm and hope, cynicism and disillusionment, and trigger linguistic responses in the form of naming, characterising, satirising and generally emoting. In addition, there are changes in the language reflecting the adoption of paparazzoid practices, exposing and judging the private deeds of public figures and bodies. At the same time, the media have imposed on themselves imperatives for ever greater immediacy and scale of news coverage, leading to ever more comment and interpretative spin to supplement actual reportage. All this is fertile ground for modem corpus linguists.

- **Methodology for dynamic corpus creation and analysis**

A particular methodological approach is appropriate for the processing of a long unbroken stretch of corpus data for MDCL. This treats the text as a chronological entity, processing it sequentially as it becomes available, and trawling through to identify changes, automatically where possible. Monitoring a single text flow on an unbroken, regular basis reveals the minutiae of innovation and change. The comparison of two corpora straddling a time gap of 30 years is a different approach to studying language change, which does not necessarily require each separate corpus to be handled chronologically, since each can be regarded as a window on a particular point in time. Comparison of two static entities will tend

to reveal long-term changes. There are details of methodology, several parameters such as the setting of time intervals for monitoring, which must be decided in the light of the particular linguistic change under investigation.

- **Tools for tracing change in text across time in dynamic corpora**

For diachronic study using two parallel corpora, software of the conventional kind is needed, involving word frequency counts, and the extraction of collocational information and concordances to various specifications. In addition, such corpora are amenable to extensive annotation, grammatical, semantic, and so on, by the appropriate taggers and parsers.

For the study of change using an evolving, chronological corpus, additional tools are required, in the form of processing software and statistical measures (Davies, unpublished) capable of recording and tracking significant change of various kinds across time. The Unit at Liverpool has such tools, and thus can automatically identify the kinds of language change outlined earlier in this paper. The AVIATOR system (Renouf 1993) monitors new words, new uses of existing words and the changing profile of the lexicon. The ACRONYM system (Renouf 1997:96-98, CollierlPacey 1997) identifies new semantic relations. The APRIL system monitors and classifies hapax neologistic word formations across time (Pacey et al. forthcoming). With these aids, linguists are in a position to describe aspects of the brachychronic (if not yet long-term diachronic)changes in text; to test hypotheses about linguistic change which are impossible to check with the naked eye; to track change and productivity in morphology, semantics, le xi co-grammar and syntax. The pedagogic, linguistic and technological applicability of such study is self-evident.

Model 4. Using the web as a linguistic resource

None of the models so far proposed quite overcomes the problem that the development of corpus resources, whether static or dynamic, is expensive and time-consuming, so that there is still no easy access to data which evidences the very rarest or very newest features of language use. The fourth model is in principle an obvious source of just such linguistic information: the web. The web is a text-based information source which has tremendous potential as a linguistic resource. It is larger than any finite corpus, constantly growing and being updated. It is broad in coverage, and potentially available to every corpus linguist without cost.

A number of corpus linguists have attempted to exploit the functionality of existing web search engines to produce contextualised information from the web in response to key words. These have typically been linguists with a historical background, who wish to trace an earlier existing word or pattern, probably found in a historical corpus, through to the present day (e.g. Bergh/SeppaenenlTrotta 1998:41-56, Brekke 2000:227-248), either to establish its continued existence, or to compare its previous meaning with its current conditions of use. They have complained of the tedium of such an undertaking.

A purpose-built facility for access to the web corpus. would be welcome. There are two main approaches currently in operation: off-line and online. The off-line approach, usually used for synchronic study, is less suited to the study of chance, since it involves, m essence, downloading a sub corpus from the web, processing it as a static entity, and then comparing it with a subsequently down loaded parallel corpus. Glossanet is one system which does this. The on-line approach, as exemplified by WebCorp at Liverpool (Renouf, forthcoming), is one which processes web contexts in real time, and which could, in principle, treat the web as a diachronic entity.

- **Glossanet**

Glossanet has been developed by Fairon (2000) to down load specific text from the web and to process it off-line according to user request. An associated tool, CorpusWeb, allows the user to download selected web-sites in corpus format, to be processed off-line on a Pc. Fairon (1999, 2000a) has implemented a drip-feed approach to identifying new words with this system, whereby daily versions of a selected web-site are separately downloaded and the contents compared with a 'filter dictionary' and other lexical sources. This is being used to update the DELA electronic dictionaries of English (FaironlCourtois 2000b), held at Laboratoire d' Automatique Documentaire et Linguistique (LADL).

- **WebCorp**

At Liverpool, our on-line linguistic retrieval system (http://www.webcorp.org.uk) has encountered fundamental problems from the point of view of monitoring change on the web. One is that the totality of the web cannot be accessed or quantified in a way that supports any standard statistical measurements of the significance of a particular change. Another is that, whilst the web is constantly growing and being updated, it is not constructed or renewed in any strict chronological sense, and its texts, whilst coded for date of installation on the web, are not coded for date of authoring, or even, in the case of published texts, of publication.

Both these obstacles may be overcome in time, but even then, the web should be viewed not as a replacement but rather as a valuable complement to the existing perfectly-honed, smaller specialised parallel corpora at Freiburg, and the open-ended text accumulation at Liverpool, and as a promising way forward for modem diachronic language study.

Model 5. Facilitating MDCL by coordinating historical and modern language resources

My fifth model for furthering the cause of MDCL would be to boost the communal store of textual data by coordinating historical and modem language resources. We need the past in order to understand the present. An amalgamation would increase the scope, timespan and continuity of resources, whilst lessening the inconvenience of having to switch from one corpus and set of tools to another.

It is also clear that the timeframe for the two fields is coming together. To historical linguists, it is moving forward. The Late Modem English period is well advanced, and the concept well established, though it is conceived of differently by individuals within that field, with references to the period as being anywhere between 1600 and 2000. Meanwhile, the timeframe for modem linguists is moving backwards. Most established modem English corpora contain text just from the last forty years or so, but specialised corpora now date back to the earlier part of the 20th century.

Interestingly, it seems that neither historical nor modem linguists have a set term for 'the English of today'. It is variously referred to by both as 'present day', 'current', and up until recently, as '20th-century'. Nevertheless, English of the 20th (and soon 21st) century is gradually becoming a focus within diachronic study for historical and modem corpus linguists alike. The historical linguistic team, Kytö et al. at Uppsala, have recently observed (Kytö et al. 2000:85), in relation to their project to create the CONCE corpus of 19th -century English, that,

> there is a scarcity of corpora covering the period immediately before Present Day English [...] A corpus of 19th -century English would thus provide researchers with the possibility of extending studies both of short-term diachronic change and of trends in Present-Day English backwards in time.

Historical linguists, particularly those associated with ICAME, are aware of modem English corpus research. Kytö et al. (2000:92) continue,

> In this respect, CONCE provides a rough 19th-century equivalent of the LOB, FLOB, Brown and Frown corpora. Studies based on these corpora have shown that a difference of 30 years is enough to study linguistic change. CONCE thus ties in with the recent scholarly interest in short-term change in diachrony.

Rissanen and Nevalainen (forthcoming) have recently conducted a diachronic study of downtoners in corpus texts from 750 to the 1990s.

The coordination of text corpora should accommodate the interrelationship between historical and regional variation which is moving centre stage. Like the Michigan Middle English initiative (McSparran 1997), it should also 'interconnect' corpora with other linguistic repositories, edited collections and bibliographies. There are currently and understandably major differences and incompatibilities between the various corpora, which must be addressed in the process of coordination. These are so many and so all-pervasive that one might be daunted. Just some that spring to mind are the problems of different standards and conventions associated with orthography, accuracy, sampling, tagging, mark-up, search programmes, storage methods; and differing restrictions regarding COpY-I right and licences. They each require political will and a great deal of effort i£1 they are to be resolved. If they can, it is already possible, in theory, to establish ı almost unbroken electronic access to samples of English text from the earliest documents to the web text of today and tomorrow.

5. Conclusion

Diachrony has not yet joined synchrony and variation as a major focus of study in modern English corpus linguistics, and MDCL is currently supported at just two research establishments: Liverpool and Freiburg. The long established work these units in the field, together with individual research efforts elsewhere, are testaments to the fact that there are so many fascinating aspects of language changer across time that can usefully be studied. A basic requirement for MDCL is the design, development and implementation of corpora, software and statistics capable of presenting and analysing the facts of the language chronologically. There are several resources, and types of infrastructure, in existence which support or be modified to support modern diachronic study, of which the most recent is the web. Modern Diachronic English Corpus Linguistics is an area ripe for growth.

References

Aarts, Jan/Pieter de Haan/Nelleke Oostdijk, eds. (1993), *English Language Corpora. Design, Analysis and Exploitation,* Papers from the 13. International Conference on English Language Research on Computerized Corpora, Nijmegen 1992, Amsterdam & Atlanta, GA: Rodopi.

Bergh, G./A. Seppaenen/J. Trotta (1998), "Language Corpora and the Internet: A Joint Linguistic Resource", in: Renouf (1998), 41-56.

Biber, Doug/Edward Finegan/Dwight Atkinson (1994), "ARCHER and its Challenges: Compiling and Exploring a Representative Corpus of Historical English Register," in: Fries/Tottie/Schneider (1994), 1-14.

Bogurayev, Branimir/James Pustejovsky (1996), *Corpus Processing for Lexical Acquisition,* Cambridge, MA & London: MIT Press.

Brekke, Magnar (1999), "When 'Empiry' Strikes Back: A Corporal Confront ation," Norwegian School of Economics, Norway.

Brekke, Magnar (2000), "From BNC to the Cybercorpus: A Quantum Leap into Chaos?" in: Kirk (2000), 227-247.

Collier, Alex/Mike Pacey (1997), "A Large-Scale Corpus System for Identifying Thesaural Relations," in: Ljung (1997), 87-100.

Davies, P. (unpublished), "Statistical Approaches to Describing Changes in Frequency over Time of Words or Linguistic Attributes," internal deliver able, APRIL project, EPSRC, Univ. Liverpool.

Fairon, Cédrick/Blandine Courtois (2000), “Extension de la couverture lexicale des dictionnaires electroniques de LADL e l’aide de GlossNet,” in *Actes du Colloque JADT 2000: 5e Journees Internationales d’Analyse Statistique des Données Textuelles*, Lausanne.

Fairon, Cedrick (1998-1999), "Parsing a Web Site as a Corpus," in: Fairon (199 1999), 450.

Fairon, Cedrick, ed. (1998-1999), *Analyse lexicale et syntaxique: Le syste INTEX, LintSV,isticae. Investigationes. Tome. XIIl (Volume special)* Amsterdam/Philadelphia: John BenJamms Publishing.

Fries, Udo/Gunnel Tottie/Peter Schneider, eds. (1994), *Creating and USing English Language c'0rpora*, Papers from the Fourteenth. International Conference on English Language Research on Computenzed COrpora Zurich 1993, Amsterdam & Atlanta, GA: Rodopi.

Granger, Sylviane, ed. (1998), *Learner English on Computer*, London & New York: Longman

Granger, Sylviane/Stephanie Petch-Tyson, eds. (forthcoming), *Extending the Scope of Corpus-Based Research: New Applications, New Challenges.* Amsterdam & Atlanta, GA: Rodopi.

Greenbaum, Sidney/Gerry Nelson (1996), "The International Corpus of English (ICE) Project," *World Englishes* 15,3-15.

Hickey, Raymond/Merja KytO,IIan Lancashire/Matti Rissanen, eds. (1997), *Tracing the Trail of Time*, Amsterdam & Atlanta, GA: Rodopi.

Hundt, Marianne (1997), "Has British English been Catching up with American English in the Past 30 Years?" in: Ljung (1997), 135-152.

Johansson, Stig, ed. (1982), *Computer Corpora in English Language Research*, Bergen: NA VF.

Kirk, John M., ed. (2000), *Corpora Galore: Analysis and Techniques in Describ ing English*, Amsterdam & Atlanta, GA: Rodopi.

Kyt6, Merja/Matti Rissanen (1995), "Language Analysis and Diachronic Corpora," in: Hickey/Kyt6/Lancashire/Rissanen (1997), 9-22.

KytO Merja/Juhani Rudanko/Erik Smitterberg (2000), "Building a Bridge between the Present and the Past: A Corpus of 19th-Century English," *ICAME Journal* 24, 85-97.

Lancashire, IanlCharles Meyer/Carol Percy, eds. (1996), *Papers from English Language Research on Computerized Corpora (ICAME* 16), Amsterdam & Atlanta, GA: Rodopi.

Lancashire, IanlT. Russon Wooldridge, eds. (1994), *Early Dictionary Databases.* Univ. of Toronto, Oct 1-8, 1993.

Lass, Roger (1980), *On Explaining Language Change*, Cambridge: CUP.

Lindquist Hans/Staffan Klintborg/Magnus Levin/Maria Estling, eds. (1998) *The Major Varieties of English*, Papers from MA VEN 97, Vaxjo 20-22 November 1997, Vaxjo Universitet.

Ljung, Magnus, ed. (1997), *Corpus-based Studies in English*, Amsterdam Atlanta, GA: Rodopi.

Mair, Christian (1997), “Corpora and the Study of the Major Varieties of English: Issues and Results,” in Lindquist/Klintborg/Levin/estling (1997), 139-158.

Mair, Christian (1997), “Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress,” in: Ljung (1997), 195-209.

McSparran, Frances et al. (1997), *The Middle English Compendium*, University of Michigan, http://ets.umdl.umich.edu/m/mec/release.html.

Nevalainen, Tertti/Matti Rissanen (forthcoming), “Fairly Pretty or Pretty fair? On On the Development and Grammaticalization of English Downtoners,” *Language Sciences*.

Pacey, Mike/ A?toinett~ Renouf/P.aul Davies/ Andrew Kehoe (forthcoming), "Monitoring Lexical Innovation across Ten Years of News Text.

Renouf, Antoinette (1993), "A Word in Time: First Findings from Dynamic Corpus Investigation in English Language Corpora: Design, Analysis and Exploitation," in: Aarts/de HaanlOostdijk 1993,279-288.

Renouf, Antoinette (1994), "Corpora and Historical Dictionaries', in: Lancashire/Russon Wooldridge (1994), 219-235.

Renouf, Antoinette (1996) "The ACRONYM Project: Discovering the Textual Thesaurus," in: Lancashire/Meyer/Percy, 171-187.

Renouf, Antoinette (forthcoming), "WebCorp: Providing a Renewable Energy Source for Corpus Linguistics," in: Granger/Petch- Tyson (forthcoming).

Renouf, Antoinette, ed. (1998), *Explorations in Corpus Linguistics*, Amsterdam & Atlanta, GA: Rodopi.

Rissanen, Matti (2000), "The World of English Historical Corpora," *Journal of English Linguistics* 28: 1, 7-20.

Sinclair, John (1982), "Reflections on Computer Corpora in English Language Research," in: Johansson (1982).