Corpora and Historical Dictionaries

Antoinette Renouf

1. Introduction

A definition of an historical dictionary would be that it is a diachronic account of a language, recording first and subsequent usages of a body of words, Of necessity, it is an account based on an observation of text, since, unlike a synchronic dictionary, it goes beyond the linguistic intuitions and experience of the lexicographers concerned.

Historical dictionaries have, until now, been manually created, That is to say, compiled by lexicographers from the citations culled manually by readers from given works. It is clear that this approach makes tremendous and even impossible demands on the humans involved, and in a recent article on the *OED* 2, Brewer (1993) has summarised her views and those of others on some of the problems involved in its production.

In the last ten years, however, the era of the computerized 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. The lexicographer can now return to source with relative ease. With the growth in computer storage capacity has come text-processing software, capable of carrying out exhaustive searches at very high speeds. This software has replaced the reader in the old lexicographic model. The lexicographer now interacts directly with the source data, producing, at the touch of the proverbial button, *all* citations for every word in a given corpus. Indeed, the headword list itself can be determined by the corpus.

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. One respected product of this type of study has been the *Collins-Cobuild English Language Dictionary* (1986). A simple modification would be to order the citations according to first and subsequent occurrence and thus allow a diachronic study within a bounded, finite, corpus, such as the works of Shakespeare.

There has recently, however, been a more fundamental development in corpus-based linguistic research, and one which I believe is of immediate relevance to historical lexicography. This has taken place in my unit, the

Research and Development Unit for English Studies. The Unit focusses primarily on current rather than earlier forms of English, but where our interests coincide with those of historical linguists is in the notion of diachronic study. What we wished to do was to record change in "English from now on", so to speak. The historical linguist, meanwhile, would like to be able to record "English up until now". But what is today's current English if not tomorrow's history? Within a project known as AVIATOR, we created a system that can search chronologically and exhaustively through a series of texts. It is this work that I would like to present in the remainder of the paper.

2. AVIATOR

AVIATOR was a three-year research project, set up by myself in the Research and Development Unit at Birmingham University in collaboration with two industrial partners, BRS Software Products Ltd and HarperCollins Publishers, and supported by the Department of Trade and Industry and the Science and Engineering Council, within the SALT (Speech and Language Technology) Programme. The funding was for three years, and the project was successfully completed in September 1993.

One major part of the A VIA TOR project concerned the development of a "dynamic corpus" processing system. By "dynamic corpus" is meant a flow of electronic textual data, such as newspaper data; in fact, the word *corpus* is somewhat misleading, in that the data was not treated as a finite or static entity, but as a changing, chronological flow. The aim was to create a series of software filters that automatically identified change in such a textual flow.

Change is identified in four ways, in terms of

- new word occurrences or "usages";
- new word uses or "senses";
- new word combinations; and
- the composition of the lexicon overall.

3. Filter 1: New Words

Filter 1 identifies new words. The procedure is as follows: the software is given a body of electronic text deemed to represent the established lexicon at the start of the monitoring process. The text could be a large static corpus, such as the 20-million-word Birmingham Corpus, or it could be an accumulation of text up to a certain date, such as several years of newspaper text; the appropriate base-line to adopt will depend on the research purpose. Filter 1 converts all the words in the text into a series of master word-lists. These master lists consist of five different classes of words: "ordinary" words, proper names, abbreviations, numerals, and queries, and they are built automatically by categorizing the words in the corpus on the basis of contextual clues, including typography. So upper- and lower-case use is looked at carefully, as are full stops, position in sentence, combinations thereof, and so on. New text then flows through the system, which matches each word against its relevant master list. Words that are not found in the master lists are recorded, with date of occurrence, and presented as new output.

So words met for the first time by our system are deemed to be "new" words; in fact they may be new, but they may instead be rare words, old words that have come back into use, or simply words which happen not to have been used in text since we began the monitoring process in 1989. Whatever they are, our system cannot fail to catch them. It presents them as potentially new words once only; thereafter, they are added to the relevant master file, with first and latest dates of occurrence registered.

In the *Times* newspapers of June 1993, Filter 1 trapped over twenty thousand potential new words, potential because many are not, or at least do not begin as, bona fide words but as errors of one kind or another. An excerpt from the "ordinary word" master file is as follows:

Word	Cumul. Freq.	First Date	Last Date
abandon	422	T27Sep89	T30Apr91
abandoned	826	T27Sep89	T30Apr91
abandoning	183	T27Sep89	T28Apr91
abandonment	115	T280ct89	T28Apr91
abandonments	2	TI4May90	TlOJan91
abandons	28	TI5Jan90	T29Apr91
abase	2	Tl5Dec90	T28Dec90
abasement	5	TO lDec90	Tl2Mar91
abashed	5	T13Nov89	T06Apr91
abate	8	T06Nov89	Tl6Mar91
abated	20	T230ct89	T26Apr91
abatement	16	T06Nov89	T02Apr91
abates	3	T02Jan91	T29Apr91
abating	16	T230ct89	T26Apr91
abattoir	12	TIIJun90	Tl6Apr91
abattoirs	9	T21May90	T05Jan91

FIGURE I: Extract from "Ordinary Word" Master List April 1991

Within the "ordinary word" category, there were over five thousand items. A sample of the output, showing something of the range of word formation that was recorded, is presented here:

3.1. *Words Formed by Conventional Means*

These are words created by the nonnal processes of word fonnation, beginning loosely with affixization, and moving on to compounding and blending. In fact, more than one type of process is evident in many cases.

Here, young people of Afro-Caribbean origin are widely seen by the law as **criminogenic**
Stephen said that the **techno-phobics** should not be put off the new generation of equipment
call Sua Emittenza (his **broadcastership**), or L'Ingegnere (the engineer)
fly out from Los Angeles to cut and **colourise** (sic) Hillary Rodham Clinton
has **conglomeratisation** been totally disastrous for British publishing
camouflaged in the sub-tropical forest, an airy **eco-creation** in wood and high-tech
related to our **tennising** boys and girls usually experience the stirring of the pride,
Increasingly, **inner-childism** is being seized on by the rich and famous to enhance their glamour
the pressures of sexual identity and patriarchal **staff-ist** attitudes
tinkle of a champagne flute cracking, the regular slither of high heel **on birthday-caked** floor.
The board agreed unanimously to provide funding for **chess-related** events
Animals do show "**pawedness**", with rats, mice, cats and dogs all having individual preferences
house with pool, built in 1900 on the outskirts of the spa **townette** of Luso
Sport will be covered, but "not in the **train-spotterish** depth of other stations"
blamed for vulgarising Althorp with all manner of parvenu **chintzery**
Wimbledon fans, undoubtedly they are middle class, either actually or **aspirationally .**
"Translations's" fault is probably that it is a bit **thinky**, but it has its compensating strengths.
"**lamontable**" - sticking to office well after the adhesive has worn out;
As these **yoblings** manifestly have no stem father of their own
My Guardian **sneaksperson** tells me that when a royal personage paid a visit there
The cats, since named Smudge and Watkins, are now residing in a **cat-friendly** household
IN your amusing dinosaur-movie **slag-off** (The Culture, last week),
only remotely cheerful plate carrier I met on the whole of my three-day **eat-over**
publicity-hungry Kusama staged orgies outside the New York stock exchange, weekly "**flesh-ins**"
President Mitterrand provoked a minor scandal by choosing the **bestubbled** designer
When the Braer foundered on the rocks at Fitful Head, predictions of **eco-doom** gushed forth
would sign up for an education degree and come out talking **edu-babble**: Piagetan theory accompanying
merchandise to buy. McDonald's will serve **dino-fries** with the bronto burgers
an array of tackily designed merchandise **dino-balloons, dino-dolls, dino-soap**
opportunity to cover her with the highly realistic contents of its vast **bronto-sinuses**.
flogging his dinosaur footage to the makers of subsequent **rex-ploitation** flicks
campaign by **profitosaurus** is pouncing on some childlike instincts
dearth of the heroic foreign correspondent **scoopsters** who populated comics in my childhood.
Gill Crabbe reports on a growing band of zealous **bat-watchers**
The prime minister approached the podium in the knowledge that a **run-of-the-processor** speech
songs teetering between ground-breaking extravagance and **anarcho-grunge** indulgence.
there was plenty of more **gawp-worthy** material.
wealth derives from the world's most successful **auto-mall**: a glamorous shopping centre for cars
At **eighty-something** she is coy about her age Dr Widdowson is far from ready to give up
those heady Summer of Love days working on the **califomication** of his brain.
Describing themselves as an **ad-hocrisy** of video punks, [they] have combined
The Desire Paths is rich ground for fans of anagrams – **fanagrams** perhaps.
Judge of his **astoundishment**, as Rikki Fulton would say, when he discovered
support for evolving "**co-opetition**", where competitors form strategic alliances

3.2. *Transliteration of Speech*

The particular subset of language that we are observing is journalism, and here a common vehicle for ironic humour is the use of actual or quasi-quotations of speech and spoken fonns.

"No sex, **purleeeze**!" That's the plea of the ambitious Hollywood exec
"Fabulous little blondes"; "**phew-wotta-scorcher** blondes"; "succulent under-25-year-olds"
the character with its unchanging response to life's perplexities ("I only **arsked…**)
Russian men do not like happiness because happiness is a little **voolgare**.
he hudnae worked fur 30 year wi' his back an he'd only hauf a **stummick**

3.3. Onomatopaeia

Cyril **whumphed** back in his seat with a sulky expression.

3.4. Slang and Miscellaneous Invention

So when I first moved to my own, personal **grot-hole** in north London
the ever enthusiastic Mr Ashdown to brush up on his **Maccers**. Was that a dagger he saw filled it with
priapic **cloggies** and beggars living in cardboard boxes

3.5. Odd Formations

Odd derivations enter the language for a number of reasons, ranging from ignorance, to an intent to amuse, or to express a new concept.

What is so funny about his "**unreverent** robes"
although "**deprival** was his element", he is all mush when it comes to gorgeous girls
galley with three banks of oars on each side yesterday braved distinctly **unjolly** boating weather

3.6. Unorthodox Spelling Variants

Oddities of spelling in the June 1993 data are not simply dismissable as errors, and are not automatically rejected at Filter 1. Time, and Filter 4, will show whether they are subsequently being assimilated into the language.

he was "**philisophically** attracted" to the lowest possible taxation
Eric Dynock **conjours** with some names from the past that bring pleasure to the present
A Svelte devotee will be spending Pounds 330 a year in the quest for **taughter** thighs
four great masses of **torturously** baroque ruins inviting the spectator to wonder
uneasy in any surroundings more **arborious** than a sparsely tubbed patio
Aorestan was imprisoned in a "**durgeon**" and his rescuer had her name shortened to "Fidelo"

3.7. Revived Words

As said previously, words that are encountered by our system for the first time are regarded as new, though they may simply be rare items or revivals or earlier usage.

they may have been done for **amuletic** purposes.
Hitchens's **macaronic** style strikes me as less dandified than dandiprat.
directed against wanton and officious **intermeddling** with the disputes of others opposed
by the **stintholders** who were local farmers with registered **stintage** rights. his prose was
sometimes "less dandified than **dandiprat**"

In the first months of monitoring, Filter 1 output will be very large, although it will lessen gradually. As the source of information for the creation of a synchronic dictionary, the output will require considerable post-editing. However, from the point of view of building a diachronic historical dictionary, where the aim is not a restricted lexicon, the exhaustiveness of this filter is a benefit.

4. Filters 2 and 3: New Word Uses or Senses

Filters 2 and 3 identify words that are being used in a new sense, or with a new reference. To identify semantic change in text by automatic means requires the selection of a linguistic criterion for newness that can be recognized by a computer. We chose to focus on the collocational environment of a word, based on our observation that this correlates with meaning in a fairly direct way. We knew that when a word begins to adopt a new collocational pattern, it is also taking on a new semantic or at least referential role. Filter 3 records new word pairings whether or not they are adjacent, within a nine word window, with a view to capturing a shift in word use across time; Filter 2 selects only those pairings that are adjacent, in a bid to discover new word combinations and nascent compounds.

As with Filter 1, Filters 2 and 3 compare each new word with its entry in a master lexicon. This lexicon, however, contains not just node words, but a record of their favoured collocates. I named this particular type of lexicon a Collocate Bank (Collier, 1993).

Words that begin to occur with a given level of significance with new collocates are presented for inspection. Statistical measures are employed to ensure that the co-occurrence of the two items is more than chance, and so only new contexts that have occurred several times are presented. In January and February 1993, some of the candidates for new usage are shown in FIGURE 2. The node word is marked "2" to indicate that it is an existing word in the Collocate Bank; its collocates are marked "4" where they are new, and "5" where they are not new but have started to occur more frequently.

2 accompanied 115
4 commander 13043.91930.0203
4 laurence 26 4 3.9837 0.0192
4 timothy 31 43.9806 0.0193
2 battle 372
4 airtours 121 54.70450.0047
2 horse 230
4 ripper IS 6 5.9718 0.0006
2 windfa\1 30
5 tax 791 14 13.55540.0000 5.5302 2 child 575
0.0001
2 algorithms 12
4 genetic 40 6 5.9961 0.0006
2 aid 406
4 wildlife 86 12 11.44480.0000
2 abuse 96
5 professional 341 4 3.8262 0.0219
1.7000 0.1463
2 ca\1er 20
4 identification 27 6 5.99560.0006
2 fish 204
4 rights 9 43.98930.0192
2 tolls 32
5 electronic 96 76.96870.0001 \.9671
0.2704
4 idea 504 4 3.9073 0.0205
2 male 310
4 midwife 24 7 6.9245 0.0001
4 midwives 20 109.9100 0.0000
2 rape 157
4 school 1046 4 3.1777 0.0391
4 causes 129 5 4.4957 0.0057
5 crime 474 11 7.58600.0000 2.0994
0.0087
4 cures 10 5 4.9583 0.0038
2 non-recourse 12
4 finance 373 5 4.9675 0.0038
2 workfare 117
4 benefit 406 5 4.6688 0.0049
4 debate 313 5 4.74200.0046
4 schemes 212 98.68190.0000
2 pancake 21
5 day 159454.76350.0045 \.6640
0.3381

FIGURE 2: Extract from Filter 3 Output from *Times* Data Jan. and Feb. 1993

The output in FIGURE 2 is not explicit, but suggestive of language change. If we now retrieve the original contexts for some of these and other items, we see more precisely what the changes are. In FIGURES 3-11, we examine some new pairings.

1 0 Mare ki\1ed by >horse ripper Mountbatten by Sean Ryan
1 0 known as the >horse ripper is caught At least
1 0 Of course the >horse ripper Rippers are far more
1 0 experts agree The >horse ripper or Rippers the police
1 0 other animals The >horse ripper he concedes is a
1 0 to understand the >horse ripper I remember a young

FIGURE 3: Extract from Filter 3 Output for *horse* plus *ripper*

This alludes to Jack the Ripper, a shadowy nineteenth-century figure said to have killed women with a knife. So *ripper* is being re-used in that established extended meaning, but in new combination with - and therefore reference to

_ horses, predominantly mares. The adjacent words constitute a new word compound.

1 0 stating it quite plainly fish Have >rights Of course
1 0 confused I mean perhaps fish Have >rights was a
1 0 But it's all true fish Have >rights is the
1 0 astounds me in this fish Have >rights business is

FIGURE 4: Extract from Filter 3 Output for *fish* plus *rights*

Here, *fish* specifies a new area of concern, an aspect of life which is or should be awarded *rights.* The two words combine in analogy with other

noun phrases containing the productive element *rights,* such as *animal rights* and *human rights.* The words, not necessarily adjacent as yet, have not settled into a compound item.

1 0 demand file on >school rape By frances gibb legal
1 0 recall a case of rape in the >school environment
1 0 a case of >school rape in their files It
1 0 of childline agreed that rape in >school was a

FIGURE 5: Extract from Filter 3 Output for *school* plus *rape*

FIGURE 5 echoes the previous figure in identifying a new area of social concern, based on analogy with existing noun phrases, such as *date rape.* In FIGURES 6 and 7 below, two new adjacent pairings can be seen, each of

which reflects early references to new, real-world phenomena in the news.

1 0 is attended by a male doctor a male >midwife
1 1 a male doctor a male >midwife and a male
1 0 male >midwife and a male partner three men It
1 1 to know what a male >midwife is about he
1 1 and jo schneider a male >midwife working in greenwich
1 1 cared for by a male >midwife could prove traumatic
1 0 to know her >midwife male or female sooner than

FIGURE 6: Extract from Filter 3 Output for *male* plus *midwife*

1 1 doctor had learnt about ambulant >epidural analgesia pain relief
1 1 successful standard rather than ambulant >epidural delivery can be
1 1 of judgment for having ambulant >epidural analgesia does not
1 1 in doctor magazine The ambulant >epidural when successful aims
1 1 pain relief in the ambulant >epidural is one of
1 1 who do choose an ambulant >epidural should be encouraged
1 1 of action of the ambulant >epidural makes this possible

FIGURE 7: Extract from Filter 3 Output for *ambulant* plus *epidural*

FIGURES 8-10 below show word pairings that are not new, but which have become prominent in early 1993. Significantly more frequent collocates of the words *tolls, alone,* and *child,* respectively, are marked by @. Sometimes these collocates Occur in environments where new collocates are also identified, as in the first lines of *electronic* plus *tolls,* giving the user a stronger sense of the change going on.

2 0 >idea of putting @electronic tolls on existing motorways and
2 0 the >idea of @electronic tolls on such routes as
1 0 ease rush hour @electronic tolls By nicholas watt february
1 0 nicholas watt february @electronic tolls are relatively simple and
1 0 @electronic versions of existing tolls cambridge city council is
1 0 asked to install @electronic tolls collect charges and use

FIGURE 8: Extract from Filter 3 Output for *tolls* plus *electronic*

1 0 @home alone for nine days while their parents spent
1 0 who helped make f@home alone the highest earning comedy
20 for @home alone having earned just from @home alone
1 0 been to see @home alone with his daughter and

FIGURE 9: Extract from Filter 3 Output for *alone* plus *home*

1 1 a veteran of other child @crime cases on merseyside
1 0 legally binding on the child issue cull @crime hartlepool
1 1 there has been a child @crime wave the signs
1 1 for the feeling that child @crime is exploding is
4 1 >causes and >cures for child @crime @Ietter february From
1 1 fraser no to tougher child @crime moves scotland By
1 1 fraser thinks scots have child @crime solution focus scotland

FIGURE 10: Extract from Filter 3 Output for *child* plus *crime*

FIGURE 11 below demonstrates a word pairing that is not new but which becomes significant at regular intervals. In the particular case, the pairing reflects an annual event, one that happens in February.

Information about shifts in word use is a vital part of a description of the changing language, whether historical or current, so the output from Filters 2 and 3 would inevitably be a useful aid in historical linguistics and lexicography. Since, however, the historical corpora that have been created hitherto are small, the statistical measures used in AVIATOR to reduce the large number of newly-contextualized words presented as candidates for new usage would not apply. Nevertheless, the system could be adapted to work on the basis of raw frequency or could even simply present every new context encountered for inspection.

1 1 as a prerequisite for pancake @day just handy with
1 1 now dreading next week's pancake @day for fear that
1 1 hayride to a clambake pancake @day requires no special
1 1 as a prerequisite for pancake @day just handy with
1 1 would probably survive the pancake @day race staged at

FIGURE 11: Extract from Filter 3 Output for *pancake* plus *day*

5. **Filter 4: Composition of the Lexicon**

Filter 4 is designed to record changes in the lexical inventory, both in terms of individual words and of the vocabulary in general, across time. It

subsumes Filter 1, making a cumulative record of the words found by that filter in monthly batches of text, together with the number of times that they occurred each time. Words enter the language at a particular time, reach a level of popularity, then settle or even disappear again, and these changes can be traced by Filter 4.

6. **Tracing the Evolution of Textual Elements to Word Status**

We mentioned earlier that many words recorded by our filters do not necessarily begin as bonafide items in the lexicon. However, some become assimilated in time. Misspellings are one case in point. There are many categories of spelling error in journalistic text, caused by mis-typing and so on, and not eliminated by such means as spelling checkers. The consequent transposition, omission, or addition of characters in a word is very common, producing such results as *socait, remeber* and *iniitial.* Some individual misspellings will, in time, become assimilated, if not officially sanctioned, and Filter 4 is designed to find out which they are. *Accomodation* looks likely to be one candidate, as shown in FIGURE 12.

Of course, this output is not in itself strong evidence of widespread usage, since it may well be peculiar to a small coterie of *Times* editors. However, it is evidence that this misspelling is now appearing in the public writing of educated writers.

1991
Apr-06: Arrange **accomodation** and tickets, and book travel separately, allowing flexibility.
May-09: Tourists were taken to alternative **accomodation** on the Dalmatian coast.
Jun-24: Their craft, a converted Formula 40 multihull with minimal **accomodation**, was
Jul-12: Tottenham's problems, being unseeded, involved the **accomodation** of an extra round
Oct-04: happily enough in converted classrooms while waiting to be given **accomodation**.
Oct-22: which effectively eliminated private rented **accomodation** at reasonable prices
1992
Jan-02: ample railway siding **accomodatlon** and direct connexion with the Southern Railway
Jan-04: **Accomodation**
Mar-Ol: the city. **Accomodation** comprises hall, lounge, kitchen, two double bedrooms
Apr-20: Pounds 10,000 a year and **accomodation** in a Pounds 120,000 rented property
Sep-19: Rights can be booked to include an **accomodation** package at various hotels.
Oct-13: a veterinary pharmacy, office and luxurious **accomodation** for the Nicholsons
Nov-29: **accomodation** and car hire are offered as a package, but are not obligatory.
Nov-29: If customers wish to book their own **accomodation** and insurance, proof of payment
Nov-29: Customers staying in privately owned **accomodation** must provide a letter confirming
Dec-12: premises of which the whole or part of the shared **accomodation** formed part,
Dec-20: It is, of course, possible to get round the **accomodation** problem. It is not compulsory

FIGURE 12: Instances of *accomodation* in *Times* Data Sept. 1989-Dec. 1993

Acronyms represent another textual element whose path into the language can be traced by Filter 4. The acronym *NIMBY,* first introduced into British politics around late 1989 by a Conservative politician, Sir Nicholas Ridley, and standing for the phrase "Not In My Back Yard", is illustrated in the following list (occurrences from *Times* Data Sept. 1989-Dec. 1993). In UK journalism, acronyms are not nowadays punctuated by full stops, so we do not see the move from *N.I.M.B. Y.* to *NIMBY* that we would have years ago with, for example, *N.A.T.O.* However, very soon, we notice the use of orthographic variants of *NIMBY,* in lower case, with and without initial capital. Soon, too, derived forms of the new word occur, in both upper and lower case. On the other hand, the parenthetical full form that accompanies new words in the language still appears several years later, indicating that users still feel that *NIMBY* is sufficiently unfamiliar to require explanation.

1990
Mar-05: concern, epitomized by the "**nimby**" response to development. If the

1991
Feb-08: context of Channel Tunnel disruption and the **Nimby** "Not In My Back Yard"
Mar-07: **nimbyism** not in my back yard. He emphasised that the South-east was
Mar-IS: Of course I'm guilty of **Nimbyism** (Not In Mother's Back Yard); and, of
May-19: (**Nimby**) syndrome in southeast England are forcing quarrying firms to seek
Jun-30: road. There are as many **Nimby** (Not in My Backyard) positions as there are
Jul-05: more recent fame to Ridley's attack on the selfishness of **Nimbys** ("Not in my
Aug-04: "You are familiar with **Nimby** Not In My Back Yard. But don't get caught by
Aug-19: feared an invasion and eventually spawned the **Nimby** (not In my back yard)
Sep-04: Ridley's contribution to English letters) **Nimby**. But it is a dictionary's
Oct-13: whose homes would be blighted had already swelled the ranks of the **Nimbys**,
Oct-22: Admittedly Cleeve Prior's stunt is a **Nimby** ("Not in my back yard")
Nov-II: Power stations usually provoke an attack of **nimby** not in my backyard. We
Dec-3I: desire to pacify the **Nimby** voters of Orpington, Stratford is the right

1992
Feb-28: pragmatism, others **Nimby** hypocrisy.
Mar-15 **Nimby**, someone in favour of development as long as it is Not In My Back Yard.
Mar-17: route. Legend has it that **Nimbys** with influence, such as Lords Palumbo and
Mar-28: fuelled by the **Nimby** factor.
May-IS: Rejecting Tory grumbles that she is just another **Nimby**, she comments: "They
May-27: New age of **nimby**; Leading Article
May-28: We have never wanted to become **Nimbys** and we sympathise with the people of
Jun-03: "invasion" ("New age of **Nimby**", May 27) would have written in such
Jun-06: of the Danube but downriver: the ultimate **Nimby**-i**sm** among nations. Austrian
Jun-19: getting it wrong. If it is in any sense a **Nimby** organisation, then it is
Jun-19: **Nimby** for all, by which I mean that there is one backyard on which we all
Jun-2S: offered a new variant of the **Nimby** (not in my backyard) syndrome: Nodam or
Jul-19: The plot is a classic **nimby** (not in my back yard) dispute in the heart of
Jul-29: **Nimby** syndrome is too emphatic. Certainly, the reaction of authority so far
Aug-06: This is a classic case of **Nimbyism** in action except, of course, that in
Aug-27: The "**Nimby**" five are characterising the USAir deal as a takeover of

Sep-15: buildings that even some hardened **Nimby-ists** might like. Higher standards
Sep-25: **Nimby**-Whistlers (not in my backyard while I still live 'ere). They are people
Sep-30: Sir, As well as Mr Russell Hawkes's **Nimby**-Whistlers (letter, September 25)
Oct-19: Mr Byham said: "This is not just a **nimby** not in my backyard reaction.
Oct-22: **NIMBYer** parts of the country. There was also the flogging by Lord Tebbit, the
Oct-31: lOft-high fence and buy a guard dog. The philosophy of **Nimby** (Not In My Back
Dec-16: developments (the "**nimby**" factor) and, nationally, when the Green party
Dec-16: part of the University of London, has come down with an acute attack of **NIMBY**

Other acronyms that have also undergone this process and are emerging as derived forms, albeit in the limited timespan of the sample, are:

Feb-17: with his youngest son, Richard Brooks, a **buppie** (black urban professional) malcontent.
Mar-03: both sides will be ordered not to cross. **Scud-busting** patrols are still being flown
Mar-21: a consumer will have to spend Pounds 5,600 on **vatable** goods before he is worse off
Mar-24: chandeliers. "It's our attempt to **de-yuppify** the place," said Derek Statt, executive

Acronyms are proliferating particularly in technical domains. In computing, the term *Rise*, standing for "Reduced Instruction Set Computer", is one, as shown below *(Times* Data Sept.1989-Dec. 1993).

1991
Apr-l1: reduced instruction set computing (**risc**) chip made by the MIPS Corporation.
Apr-11: So far **risc** chips have been used mainly for workstations extra-powerful
Apr-11: conform with it. IBM and Sun are already established with **risc**-based
Apr-11: new technology before IBM. Compaq expects to produce a **risc**-based machine
Apr-23: known as **risc** (reduced instruction set computer). "The new machines may be
Oct-03: Apple will adopt IBM's reduced instruction set computing (**risc**)
Oct-03: applications from both companies on **risc**-based hardware.
1992
Jan-10: field of **RlSC** (reduced instruction set computing) chips, a new generation of
Jan-29: computing (**risc**) system, considerably enhancing computing power.
Feb-07: that goes by the name of **risc**, or reduced instruction set computing, a
Feb-07: ull has failed to develop its own **risc** technology, and has instead opted for
Feb-07: a **risc** technology based on the widely available Mips chip, which is supported
Feb-07: Hewlett's proprietary **risc** technology is more "up-market" than that of Mips
Feb-07: holds that Hewlett's **risc** technology is superior to IBM's. Hewlett has
Feb-07: instruction set computing (**risc**).
Feb-07: Microsystems, over which **risc** systems will become the industry standard.
Feb-28: reduced instruction et computing **risc** which speeds up a processor by
Feb-28: sing alpha, Digital plans to rework its entire computer line with **risc**,
Apr-07: Intel chips and moving through the 286 and 386 to 486. Faster **RlSC** -based
Apr-21: **RlSC** microprocessor.
May-08: They use a technique known as reduced instruction set computing (**risc**),
May-08: involved a workstation with several **risc** chips in the processor rather than
May-15: **risc** game;Online; Infotech Times
May-15: Power PC single-chip **risc** microprocessors.
May-29: The computers will use powerful **risc** processors, which Apple says will Jun-27: decision by Olivetti to adopt DEC's Alpha **RlSC** (reduced instruction set
Jun-27: computers. **RlSC** chips are faster than the present type. The choice of an
Jun-27: appropriate **RlSC** technology was also the main consideration in Bull's
Jun-27: **RlSC**, with the market leaders, such as IBM, Hewlett-Packard and DEC, trying
Sep-13: Then there is **risc** (Reduced Instruction Set Computers), an alternative
Sep-13: form the circuits on the silicon, to make the system run fast. A **risc** design
Sep-13: But **risc** technology has one big drawback. Because it operates on a new Sep-13: instruction set, it can only run new, specially-written **risc** software. Complex
Sep-13: The **risc** challenge has, however, forced Intel to make a number of
Sep-13: chief executive, has confidently declared that the war with **risc** is over.
Sep-13: "The 486 incorporated **risc** technology, and the P5 (the next Intel chip)
Sep-13: Intel's Cisc design, because it runs slower than its rivals' **risc** design,
Sep-13: will have to develop new generations of chips much faster than the **risc**
Sep-13: the surface of a single chip. The **risc** manufacturers reckon they can reach

RISC is evolving slowly as a word. It displays, at least in the sample offered above, relatively little of the linguistic flexibility potentially available to it.

There are no derivations of the *RIScy* variety, and no idiomatic play of the *Take a RIse* kind, as would be expected in newspaper headlines or in adver tising copy. It seems unlikely, however, that computing specialists do not indulge in such linguistic activity.

7. **Tracing the Fortunes of a Word**

When a new word is introduced into the language, it is commonly marked in some way. In its early mentions, the signals can take the form of inverted commas, as in

Mar-22: missed the mayhem generated by the "**hovver bovver**" of recent years as Qualcast

or, in a similar fashion to the treatment of new acronyms mentioned above, by appositional phrases functioning as definitions, such as

Feb-17: at any rate." The result is an **allobiography** (about other people rather than oneself)

After some time, when a word or a phrase becomes popular, it often also becomes the target for popular allusion by analogy. *Watergate* was a famous example which to this day is still being exploited, as in the latest *Whitewater-gate*. The word *Thirtysomething* was coined in an American televised drama series that came to Britain in the 80's. *Times* data reveals several allusions for "Thirty something" variants (Oct.1990-Mar.1991):

Oct-22: Newcastle's greying defenders, the "**thirtysomethings**" of the division, are vulnerable
Jan-06: endlessly patronised by the forty- or **fiftysomethings**. The reality
Jan-20: Edmond) for details. Robert Cray, who at **fortysomething** can just lay claim to
Jan-07: America's baby-boom generation, now **thirty-something** or slightly older with young
Feb-17: She magazine, the post-feminist glossy for the **thirtysomething** working mother
Feb-17: just been fired from her Guardian column called **flftysomething**, Brown snipped back,
Sep-26: shop-floor workers, from the school-leaver to the **fiftysomething** of both sexes,
Feb-24: The three rather disparate forty-somethings likely to dominate the new bidding
Mar-06: New York _ ALL got up to look **forty-something**, Tracey Ullman imitates her 17-year

Mar-24: Dream as interpreted by the pampered **twentysomethings** of New York society".
Sep-03: the housing ladder. While **twentysomething** Americans and Continental Europeans
May-21: the number of people in the forty-, fifty-, and **sixty-something** generations and beyond
Jun-16: Kennedy's working-class hero image and the **fortysomethings** will come creeping out
Jul-12: with her eyes at street level – **Fortysomething** come in two types: those whose popular

Thirtysomething was an example of the kind of labelling of social types that is very popular in British and American English. In Britain, we also have *buppies, yuppies, dinkies, Essex Man,* and *Essex Girl,* as other recent cases.

Thirtysomething opened the door to a characterization of people of a shared age range that implied a correspondingly shared lifestyle and preoccupations, as the word *teenager* had done, and it obviously met a need. The word *grunge* met another need, as can be seen in FIGURE 13. Already extant in the spoken, semionomatopoeic form *grungy,* used to characterize nasty, dirty or unkempt things, it became confused with the similar item *gunge,* which was used of anything which, largely through neglect, had become dirty and often malfunctioning (such as machinery) and so was *all gunged up.* Out of thisemerged a new use, a specific reference to a kind of music that was so named by association with young musicians who adopted a certain style. It is clear from the extract that *grunge* went on to fulfil a further need, that of naming the fashions and lifestyle that are a parody of the ones informally identified by *gunge* and *grunge* in their earlier incarnations. In its new role, *grunge* has moved firmly into the sphere of written language. Filter 4 will monitor its future path and show whether it is as ephemeral as the fashion it names, or whether, when and how it will settle into the lexicon.

The changing frequency profile of a word, or indeed of several associated words, can be accessed by the user in graphic form that is easier to interpret. In appendices I and 2, the fortunes of some words with variable spelling in the Helsinki Corpus (Rissanen, Kyto, and Palander-Collin, 1993) are traced: Appendix 1 deals with three variants of the word *thing: ding* (where *d* represents *thorn), thyng* and *thing;* Appendix 2 compares the variants *nyce* and *nice.* It is assumed that this type of graphical information will be of interest to historical linguists for the clear overview that it can give.

8. **Conclusion**

The application of the A VIA TOR system to historical text will bring all the advantages of modern, computerized lexicology to the field of historical linguistics. Increased speed will be of major benefit: until now, updates of manually compiled historical entries have taken place about every hundred years or so, but with a computer-aided system like this, updating could be

1990
Oct-22: flirtation with the primal **grunge** of noise, only they know.
1992
Feb-23: those **grunge**-rock decibel junkies Tin Machine, David Bowie has begun work on
Mar-15: greatly preferred the **grungy**, reverberative efforts of those pioneer
Mar-22: Burton Road, to be greeted by an incredibly **grungey** man, moodily complaining
Apr-OS: **grunge**-rockers Dinosaur Jnr, were just a bit too fast and frantic for the
Apr-09: high "**grunge**" factor, was sufficient to compensate for any lack of detail.
Apr-11: sauce that bears no relation to the genuine article: ointment plus **grunge**.
Apr-25: Seattle takes the **grunge**; Rock
Apr-2S: had shown the makings of a promising cult band. Like their **grunge** rock
Apr-25: Although **grunge** shares many superficial characteristics with heavy metal and
Apr-25: have finally brought **grunge** to the masses.
Apr-25: On numbers such as "Rude", "Distracted" and "Murder" (**grunge** bands
Apr-2S: Several all-girl or girl-led groups operating in the **grunge**/noise idiom have
May-23: performances brought to you by the denizens of Seattle, "the **grunge** capital
May-23: swiftly subsumed by a swirling surge of pre-**grunge** guitar riffing. From then
Jul-26: **grungy** bedsit; we learn the horrors of Ella/Max's life as a crane driver
Aug-01: and are starting to sound like a conventional post-**grunge** rock band. Still,
Aug-02: calls for a certain **grungey**, amateur quality which might be squashed by
Aug-30: be found trotting to The Globe the Groucho Club of the **grunge** set, favoured Sep-02: that inspired the coinage "**Grunge**" was present in fact but not spirit.
Sep-11: other rock band. Their elephantine **grunge**-metal not only failed to leap the
Sep-13: **grunge** rockers Husker Du have concocted some delicious new wine in old
Sep-23: the rise and rise of **Grunge**-Rock band Nirvana. They have injected the old
Oct-10: rend the air with some ear-splitting **grunge**-rock.
Oct-10: world, independent and, well thrash metal **grungers**.
Oct-22: melody with modem attitude and a modish touch of **grunge**, it seemed, for
Oct-25: charts with all that **grungy** metal and un lyrical grooving. Female voices now
Oct-25 version of that sort of guitar-heavy **grunge** associated with the mighty
Oct-25: Making a choice of **grunge** or sleaze; Music; Scotland
Oct-25: difference between **grunge**, industrial, sleaze, hardcore, grindcore, or death
Oct-25: meets **grunge**. **Grunge** is the one where you really cannot hear the words at
Oct-25: Gruntruk is industrial **grunge** a bit like Nirvana or Mudhoney. The band's
Nov-08: We have "**grunge**" things happening in America right now. People make a style
Nov-08: New York seduced by the tramp; **Grunge**; Fashion
Nov-08: THE word on the street in New York this winter is **grunge**.
Nov-08: **Grunge** is aggressively casual: jeans worn back to front, so big that they
Nov-08: collection that was a homage to **grunge**.
Nov-08: Another designer making his big statement with **grunge** is Christian Francis
Nov-08: **grunge** stories abound, like the one about the powerful, New York magazine art Nov-II: of the look is GruDge, a supposedly anti-fashion movement that mixes up hippy
Nov- 11 **Grunge** comes out of Seattle and affects a certain wildness of behaviour,
Nov-15: **Grunge**; Nasty cult of the nineties
Nov-15: **Grunge** half grotty, half grunge, a strange amalgam of street fashion and
Nov-15: time. Bob Geldof was unwittingly **grungy** at birth; so were the actresses Julia
Nov-15: **Grunge**-designer Helen Storey's London shop, looking at her latest collection
Nov-15: winter, is the first **Grunge** movie, starring Matt Dillon.
Nov-15: In **Grunge** speak, awesome equals good, harsh is hard; shine that scene

FIGURE 13: Some Instances of *grunge* in *Times* Data Qct. 1990-Nov. 15, 1992

every few years. If exhaustive accounts are desired, thoroughness and completeness can also be assured, since the computer cannot overlook things that the human eye can miss. If not, objective criteria for selecting a restricted lexicon are also available.

The AVIATOR system requires a dynamic text flow, but it can be applied to collections of historical text provided that they are marked for chronology, and can therefore be treated as though they were dynamic. Some modification of the software will be required, to allow for the particular characteristics of historical text, such as diacritics and variant spelling, but this is feasible. Another constraint in the application of the system to historical text lies in the small amount which is currently available in electronic form. There exist a range of smallish, carefully-honed corpora such as the Helsinki Corpus. These are typically indexed and well-documented, so that the information provided by our Filter 1 can be extracted by other means. On the other hand, information about change in word use provided by Filters 2 and 3 would be of great value. Although the small corpora do not at the moment furnish sufficient text for a change in usage, or meaning, to be reliably identified by the normal statistical means, the system could be modified simply to draw attention to every instance of a new collocational pairing, or to make minor reductions in otherwise bulky output by taking raw frequency into account. Meanwhile, information of the type provided by Filter 4, about changes in the path of a word or words, would be of interest and could be extracted even from a small corpus.

Antoinette RENOUF
Research and Development Unit for English Studies
University of Liverpool, England

Bibliography

BREWER, Charlotte (1993). "The Second Edition of *The Oxford English Dictionary", The Review of English Studies,* NS 44.175 (August): 313-42.

COLLIER, Alex (1993). "Issues of Large-Scale Collocational Analysis", *English Language Corpora: Design, Analysis and Exploitation* (ed. Pieter De Haan, N. Oostdijk & J. Aarts). Amsterdam: Rodopi.

RENOUF, Antoinette (1987). "Corpus Development", *Looking Up* (ed. John Sinclair). London: Collins ELT.

RISSANEN, Matti, Merja KYTO & Minna PALANDER-COLLIN, eds. (1993). *Early English in the Computer Age: Exploration through the Helsinki Corpus.* Topics in English Linguistics, 11. Berlin and New York: Mouton de Gruyter.