Ed. Mattioli 1885 (2020) ISBN 978-88-6261-777-2

MODERN DIACHRONIC CORPUS-ASSISTED LANGUAGE STUDIES: METHODOLOGIES FOR TRACKING LANGUAGE CHANGE OVER RECENT TIME

Glen Michael Alessi[1], *Alan Partington*[2]
[1]University of Modena and Reggio Emilia, Italy; [2]Bologna University (Forli), Italy

Ed. Mattioli 1885 (2020) ISBN 978-88-6261-777-2

Modern diachronic corpus-assisted language studies: methodologies for tracking language change over recent time

Glen Michael Alessi[1], *Alan Partington*[2]
[1]University of Modena and Reggio Emilia, Italy; [2]Bologna University (Forlì), Italy

Abstract. This paper presents a description of the tools and methodologies employed in the novel discipline of modern diachronic corpus-assisted language studies. The main instruments are a set of three 'sister' corpora of parallel structure and content from different moments of contemporary time, namely 1993, 2005 and 2010, along with a number of corpus interrogation tools. The methodologies are the particular techniques devised by the SiBol research team[1] for employing these interrogation tools to shed light on the various research questions treated in the paper. The first part of the paper outlines ways in which these tools and techniques can be used to track changes in the grammar, lexis and discourse practices of UK broadsheet or 'quality' newspapers. Given the important role of newspapers, some of these changes may well be indicative of general changes in UK written English. The second part, instead, describes a number of studies conducted by the research group into how the reporting of various social and cultural themes and issues, ranging from what is seen as a *moral* issue, to the rhetoric of appeals to *science*, to how antisemitism is debated, has developed over the time period in question. The concluding section discusses the relationship between the methodologies employed in modern diachronic corpus-assisted language studies and wider scientific research methodology.

1 SiBol is a portmanteau of Siena and Bologna, the two universities involved in initiating the project. http://www3.lingue.unibo.it/clb/

1. Contemporary language change

1.1 Historical linguistics

Historical linguistics and the study of language change began to develop in earnest during the nineteenth century, largely as the result of a relative shift of interest from the high-prestige, literary but dead languages of antiquity to living languages, in which variation over time could be viewed as an ongoing process, not one which had reached a supposed conclusion. Such change is, as Lyons put it, 'universal, continuous and […] regular' (1981: 179) (although, as we shall see, the last two require qualification). Much of this early work was carried out in German by the Neogrammarian (*Junggrammatiker*) school and a great deal was retrospective and classificatory, that is, concerned with tracing how modern languages had developed from earlier phases ('parent' or 'proto-' languages) and aiming to group languages into 'family trees' by studying resemblances among them. A good deal of attention, however, was also paid to identifying potential mechanisms of language change. Of particular interest here is the wave theory (*Wellentheorie*) propagation of change, that is, that a particular form may ripple outwards from a centre outwards to a larger community and replace previous usage. This is often explained in terms of social prestige, that is, if the incoming form is associated with a high status group it has a considerable chance of success. The extended *of* genitive, for example, spread throughout English dialects due to the influence of high status of French / Latin after the Norman Conquest (Fries 1940). More recently Labov ([1966], 2006) showed how younger speakers in New

York were more likely to use post-vocalic [r] in words like *fourth* than older speakers, given the higher status of rhotic accents in the US, precisely the opposite, we might note, of the development of English in England, where the disappearance of post-vocalic [r] began amongst the upper-classes of southern England (Wells, 1982). However, we need a careful definition of *prestige*. Not all forms which extend through contemporary English originate in economically advantaged groups, as the spread of numerous terms from young Afro-American argot testify. We might talk more generally of an 'attractiveness principle' in language innovation.

Compared with the very impressive body of work on reconstructing language change throughout history, relatively little consideration has been given until comparatively recently to tracking *contemporary*, that is, very recent language change over a brief time period due to two interlocking factors; firstly, the difficulties in designing suitable tools and methods for doing so and a conviction in some quarters that, as Bolinger puts it: 'change is seldom on a noticeable scale … [o]ur failure to see the stirrings going on around is due to the brief sampling of time that even the longest human life encompasses' (1975: 385).

However, corpus linguistics has very recently begun to deliver new instruments to study developments in contemporary English as well as the methodological techniques to employ these tools. Moreover, it may well be that today is a particularly interesting time to study language change. Although as mentioned above, Lyons noted that change is regular and continuous, he also adds that '[l]anguages change more rapidly in certain periods than they do in others' (1981: 183). It seems reasonable to assume that, given both its increasing role as a global lingua franca and the influx of new technological vocabulary (1.4 below), the English language is currently in a phase of rapid transformation. It might also be hypothesised that this change is likely to accelerate. It may be possible to devise ways to test these hypotheses using language corpora.

In addition, a relative paucity of data, that is, of surviving texts, in most historical periods has meant that it is often impossible in practical terms to pay much attention to differences among genres or text-types (though a great deal of attention has been paid to differences among dialects). The comparative ease of access to and of convertibility into machine-readable form of a wealth of texts of various types in the internet age means that, using corpus linguistics techniques, we can study the ways different text-types may evolve over brief periods of time. In this paper we concentrate on the language of UK broadsheet newspapers.

1.2 Corpus-assisted studies of contemporary change

The first systematic use of corpora for the study of contemporary language change was the body of work by Mair et al who conducted, mainly at Freiburg University, several studies comparing the language contained in the LOB (Lancaster-Oslo-Bergen) corpus of UK English in the early 1960s and the sister FLOB (Freiburg-Lancaster-Oslo-Bergen) corpus of UK English of the early 1990s, which they constructed themselves specifically to mirror LOB and facilitate diachronic comparisons. Each corpus consists of 500 extracts from a wide range of different written discourse types of approximately 2,000 words each, giving a total size of one million words per corpus.

Using such comparatively small corpora, Mair and his co-workers were able to conduct studies of changes in the behaviour of very frequent grammatical words or constructions, for instance, the process of modalisation of the word *help* (Mair, 1995), changing patterns of verb complementation (Mair 2002) and changes in parts-of-speech frequencies (Mair et al 2002). Baker (2009) has compiled a third LOB-type corpus of English from the mid 2000s to enable studies of more recent language developments. However, some more recent corpora, such as the 400-million word Corpus of Historical American English (COHA) and the 400-million Corpus of Contemporary American English, 1990-2012 (COCA; Davies 2009; 2010), both of which contain texts which can be systematically searched by date, and the SiBol project (see next paragraph) have at their disposal much larger corpora, which means that studies of less frequent grammatical structures and also of lexical – as opposed to grammatical - items also becomes feasible. The employment of very large corpora also opens up entirely new avenues of research in modern diachronic linguistics; we can study meaning change, especially of sets of related

lexical items, in relationship to both internal linguistic factors and also in response to external social influences. Being in a position to study lexical patterns and how they evolve over time we are also able to study changes in discourse processes and how these might relate to more general language change. In this paper, for instance, we discuss how UK 'quality' newspapers appear to be adopting some of the language practices once thought typical of their downmarket counterparts, the tabloid papers, and the role this might play in propagating more permanent language change. Finally, as will see in section 2 we can also track recent developments in social cultural and political attitudes as reflected in the newspaper data.

The SiBol project has employed mainly but not exclusively the Siena-Bologna Modern Diachronic Corpora (the SiBol corpora), which are three sister corpora of UK newspaper texts from different but contemporary periods in time, designed and compiled to be as alike as possible to eliminate potentially complicating variables. They contain all the articles published by the three main UK broadsheet or so-called 'quality' newspapers, namely *The Times*, the *Telegraph* and the *Guardian* in the years 1993 (the SiBol 93 corpus), 2005 (the SiBol 05 corpus) and 2010. They contain, respectively, circa 100 million words, 150 million and 140 million words.

The main software employed for the grammatical investigations described in this first section is the *WordSmith Keyword* tool, one of the *WordSmith Tools* (version 5.0) suite of programs (Scott 2008). This allows us to compare the *relative* frequency of items in any corpus with reference to another corpus. The analyst first prepares a list of the items in the first corpus, known as the target corpus, in order of their absolute frequency, using the *Wordlist* tool. The same procedure is followed for the second corpus, known as the reference corpus. The *Keywords* tool can then compare the contents of the two lists and those items which occur statistically significantly (using chi-squared or log-likelihood tests) more frequently in the first list are themselves put in an ordered list. The more statistically significant the item, the more *key* it is, the higher it is placed on the list. This keyword list, providing an ordered series of items which are *salient* in one corpus compared to another corpus, is likely to suggest items which warrant further investigation. The procedure can then be repeated but by inverting the two corpora to reveal the items which are salient in the second corpus. Following this methodology, then, three lists of keywords were produced, one of the salient items in 2010 newspapers relative to 1993, one of the 2005 newspapers relative to 1993 and the third of the key items from the 1993 newspapers relative to 2010 and 2005 combined. Even when setting the *WordSmith Keyword* statistical significance setting at the most rigorous level envisaged (that is by setting the *lowest* p-value available, namely p = 10^15), the two corpora being compared were so large that each list contains over 7,000 items. However, for practical purposes the first 2,500 items in each list will be taken into consideration.

These keyword lists were then examined for items and, especially, sets of items which might be of interest in the context of the particular piece of research in hand. In the following section, we will give an overview of those sets which might offer clues about changes in UK newspaper language, both grammatical changes and vocabulary developments. In section 2 of the paper, we will instead look at a number of corpus-assisted studies which have been conducted with the aim of tracking discourses around social, cultural and political issues over this recent time period.

1.3 Systemic changes: grammar and orthography

One interesting feature of the comparison of the 2005 / 2010 data with that of 1993 are distinct differences in the relative frequency of a number of modals, prepositions and linkers. As regards modals, in the more recent lists we find *can*, *can't*, *cannot* and *need*, whereas in the 1993 list we find *would, shall*, *should* and *may*. Clearly the former items have not suddenly entered or the latter suddenly left the language, but we can witness a shift in their popularity in newspaper prose.

Regarding prepositional use, in 1993 we find as keywords *against*, *under*, *between*, *upon*, *throughout*, a*mong*, *within*, *towards* and a*mongst*, whilst in 2005 and 2010 we come across *with*, *across*, *alongside*, *below*, *onto* and *around*. Some of the 1993 words may have fallen away because of their relatively formal air - notably *throughout* and *amongst* - but it is not easy to ex-

plain the distinct changes in popularity of most of the others.

Linkers in the 2005 / 2010 keyword lists include *but*, *also*, *because*, *then*, *while* and *alongside*, whilst those the 1993 list contains *thus*, *therefore*, *moreover*, *nevertheless*, *in spite of*, *whilst* and *indeed*. There would appear to be a distinct movement over time towards the use of less formality in cohesive expressions.

A number of previous corpus-based studies have noted an increased use of the progressive verb aspect. Aarts, Close & Wallis (2010) note that the progressive is being used more often in spoken interactions, whilst Leech (2004), Leech & Smith (2006) and Mair & Leech (2006) include the increased use of the progressive among their changes apparently indicative of 'a suspected trend of 'colloquialization" in writing (Leech 2004: 63). The keyword lists of the more recent data include, amongst others, *going*, *getting*, *looking*, *doing*, *playing*, *drinking*, *thinking*, *watching*, *working*, *having*, *using*, *eating*, *dancing* and *wearing*. There are no equivalent *–ing* verb forms among the top keywords for 1993. Leech & Smith (2006) also claim that questions of all kinds are indicative of colloquialisation, and the 2005 list indeed includes *where*, *when*, *why*, *how* and *what*.

Conspicuous among the 2005 and 2010 keywords are first and second-person pronouns. High in the 2010 list for example we find (in decreasing order of keyness): *your*, *you*, *I*, *we*, *my*, *me*. Equally prominent are pronouns with verb contraction, for example (again in order): *I'm*, *we're*, *I've*, *we've*, *you'll* and *I'd*. We are witnessing an increasing familiarisation in the prose style and writer's stance towards the reader in UK newspaper language.

Contracted forms in general abound in the 2005/2010 keywords including: *it's*, *he's*, *there's*, *she's*, along with a large variety of negative contractions such as *can't*, *don't*, *didn't*, *doesn't*, *wasn't*, *isn't*, *won't*, *couldn't*, *wouldn't*, *aren't*, *haven't*, *hasn't* and *weren't*. It can be argued that this is not a grammatical change but simply an orthographic one. Nevertheless it is a significant event, especially were it found to be repeated across a large number of text-types. Both ontogenetically and phylogenetically, speech precedes writing and 'conversation is the most commonplace, everyday variety of language, from which, if anything, the written variety […] is to be regarded as a departure' (Biber et al 1999: 1038). The increased use of contractions is evidence that writers feel a need to reduce the distance, the departure, from spoken language.

In fact the majority of the developments discussed above are indications of an increase in the personalisation or familiarisation of newspaper register over the thirteen years between the two corpora. This is perhaps an unsurprising finding. Fairclough has written on what he terms the *conversationalisation* of media discourse (1995). As regards newspapers in particular, McNair (2003) describes what he calls the *tabloidisation* of UK so-called quality newspapers, whilst Carter & McCarthy note two particular linguistic features used to attempt this familiarisation with the reader:

> journalists also achieve impact and get on a 'conversational' wavelength with their readers by using common spoken discourse markers and purposefully vague language in a projected conversational exchange
> (Carter & McCarthy 2006: 238)

As yet further evidence of the trend of familiarisation we note that the 1993 keyword list contains a considerable number of formal terms of address or personal appellation, all of which disappear from the 2005/2010 lists. These include *Mr*, *Mrs*, *Lord*, *Dr*, *Sir*, *Lady*, *Rev*, *Herr*, *Signor* and even *President*. The UK press has curtailed its use of formal courtesy forms.

1.4 Changes in lexis

New words enter a language all the time, in fact '[a] language grows in the number of its words as the societies that use it create new entities that have to be named' (Bolinger 1975: 384). Given the increasing rate of technical innovation, it is no surprise to find a very large number of new-entity naming terms among the 2005 versus 1993 keywords, including *internet*, *website*, *email*, *DVD*, *mobile* (phone), *download* and *blog*. But even comparing the 2005 and 2010 frequency we come across such new entries as *tweet*(*s*) / *tweeted*, *scrappage*, *networking*, *smartphone*(*s*), *ipads* and *kindle*. We also find familiar items acquiring drastically new senses, for example *drone* indicating a form of unmanned aeroplane or *android*, a kind of telephone. Sometimes this

process can happen extremely quickly. The item *contagion* occurs 58 times in 1993, 56 times in the 2005 data but 448 times in the 2010 corpus, entirely due to its new sense as *financial* contagion among markets.

Duguid (2010a) notes a number of new slang items among the 2005 keywords including *chav*, *bling*, *geeky*, *feisty*, *rapper* and *so … not* (as in 'You are so not allowed in the front room', SiBol 05). Not all slang items are new forms however, once again they are frequently pre-existing forms which have acquired new argot senses. She notes how *edgy*, from meaning 'tense' is often now used by a younger generation to mean 'cutting-edge', 'challenging the norms' (2010a: 117), and how *skinny* is no longer always used as a pejorative meaning 'too thin', but can be favourable as in *skinny jeans* and in describing the fashionable *skinny latte*. These are both instances of semantic amelioration, long recognised as an aspect of language change. One of the main processes of both amelioration and pejoration consists in a word developing a new meaning alongside the existing one and, over time, especially when the old and new senses are in evaluative conflict (that is, one positive and one negative) it is often the case that one of the two falls into disuse.[2] If it is the old sense then we have an example of amelioration or pejoration. It is too early to tell whether the newer more positive sense of *edgy* adopted by younger speakers will – by the 'attractiveness principle' mentioned in 1.1 - eventually force the older one into obsolescence.

Beyond this and more in general, Duguid notes how UK papers now employ terms of strong evaluation, usually positive evaluation, far more profusely than in 1993, whether such terms be semi-slang, such as *iconic*, *pivotal* and *vibrant* or mainstream items such as *incredible*, *amazing* and *terrific*. She concludes that such usage is another indication of the familiarisation of UK 'quality' newspaper prose we also witnessed above and that 'the nature of "quality" broadsheet language has changed considerably, that it has adopted the use of less measured language and has become more imitative of the orality and informal language that used to be characteristic of the tabloids' (2010a: 135).

Comparing diachronic sister corpora is also a highly effective way of identifying another aspect of vocabulary change, namely, what Lyons amongst others calls *codability*, the degree to which a concept is lexicalised as a term by a language community. This is very frequently achieved by the compounding of two words to compose a single lexical item, often transitioning through a phase of hyphenation: 'the separation or lack of it in writing is a fair indication of how deep the heat of fusion has penetrated, how much the individual component has kept of its own identity' (Bolinger 1975: 111). Between 1993 and 2010 the following have all followed a path from separation (either as two words or hyphenation) to fusion in the UK press: *shortcuts*, *ongoing*, *wellbeing*, *highlights*, *turnout*, *makeover*, *longterm*, *meltdown* and *outcome*(*s*). As with meaning change, fusion due to familiarisation can occur with great speed. The item *bailout* occurred only nine times in 1993, compared to 29 occurrences of *bail-out*. Even in 2005 occurrences of the two were not uneven, 45 and 38 respectively. But in 2010, *bailout* occurs 1,517 times compared to 1,041 occurrences of the hyphenated form; the term has become both much more frequent and liable to fusion thanks to the intervening banking crisis.

It may be objected that some of the developments described above, for example, the increased use of evaluative language and of slang, are evidence not so much of genuine systemic language change as of the language practices of UK print journalists. It might further be argued that even what are undoubtedly language changes are only changes in *newspaper* language, not in the written language 'as a whole'. However, we need to bear in mind several factors. Firstly, newspapers are not a single linguistic genre but comprise several text-types, including reports, comment, letters, and so on. Secondly that, as mentioned in 1.3, similar changes in discourse practices in several other text-types have been described by other authors. Thirdly, newspapers hold an influential linguistic role. The texts they contain are some of the most widely-read long texts in society, either in newsprint or internet versions, especially among the educated – and influential

2 One important exception to this process is when the differently evaluated senses exist in different text-types or fields of discourse. A good example is the item *orchestrate*. In news and politics in co-occurs regularly with negative items such as *violence*, *attacks*, *threats* and so on, whereas in sports, to *orchestrate a move*, *an attack* or *a team* is a good thing (Morley and Partington 2009). Neither sense, at least from the evidence in the SiBol copora, shows any sign of disappearing.

– sections of the population. The language they contain both reflects changes in the surrounding society including in its linguistic practices but it also acts as an important vehicle for spreading change. The wave theory of change mentioned earlier posits the existence of both centres where change occurs and mechanisms for propagating the wave of change. Newspapers are both, they are foci where grammatical, orthographic and semantic change can take hold and they constitute a means of communicating new forms, senses and uses to a wide audience. Finally, it is possible to overstate the distinction between changes in practice and language change. Language *is* the way it is used, and changes in practice can, in the long term, lead to systemic change, for example, a decline in the use of an item can eventually lead to obsolescence, to its falling out of the language, whereas an increase in use of a word often results in its developing new meanings. Similarly, orthographic changes which catch on via the attractiveness principle become permanent (who, today, would write *myself* as two words or *inasmuch as* as four words rather than two?). Moreover, evolutions in language practice are an interesting topic of study in themselves; they still require explanation.

2. Developments over time in the reporting of social, political and cultural issues

In this second part we will describe how the use of a set of modern diachronic corpora like the SiBol suite can reveal ways in which the behaviours - and therefore meanings - of various lexical items used in the UK broadsheet press relating to particular social, political and cultural issues and attitudes have changed in subtle ways even over brief periods of time.

For these types of studies, alongside the frequency word-lists and the keywords lists, considerable use is made of the concordancer tool. The concordancer extracts as many examples as the analyst wishes of the word or expression under analysis - usually known as the search-word or search-item or node - and arranges them in a concordance, that is, a list of unconnected lines of text that have been summoned by the concordance program from a computerised corpus, with the node located at the centre of each line. The rest of each line contains the immediate co-text to the left and right of the node. It is generally possible to specify the number of characters of co-text from around, say, 40 to, realistically, around 600 on each side. For example:

1 in stanley's exploration of images **fraught with** a sense of millennial angst, but
2 tles are as enigmatic as they are **fraught with** a bemused paranoia: I Was Overcom
3 f the work, one of gay detachment **fraught with** a sense of destiny, as is everyth
4 , the reality of being a child is **fraught with** absurdities. Children are the onl
5 ment of ways forward in a society **fraught with** alarm and confusion over unruly y
6 hasing a property overseas can be **fraught with** all kinds of problems. Look for a
7 ur total reliance on computers is **fraught with** all kinds of dangers. 31 July 200
8 n gnome was a totem of our times, **fraught with** all kinds of symbolism: economic
9 he whole business of nicknames is **fraught with** ambiguity. At their best, nicknam
10 e baby-naming process can also be **fraught with** anguish. Catherine, 32, a design
11 er personal experience, are often **fraught with** animosity and conflict. It's a pi
12 is Davis Cup debut on an occasion **fraught with** anxiety, not least because politi
13 is Davis Cup debut on an occasion **fraught with** anxiety, not least because politi
14 ture husband. Our wedding day was **fraught with** anxiety, my mother saying she was
15 ensive purchases and decisions is **fraught with** anxiety. No wonder so many of us
16 dustry, blighted by urban sprawl, **fraught with** appalling social problems? Or was
17 hird Reich). The 20th century was **fraught with** atrocity. The atomic holocaust of

Figure 1: A concordance of the expressions *fraught with* from a corpus of UK newspapers.

Such a list enables the analyst to look for eventual patterns in the surrounding co-text, which proffer clues to the use of the node item. In the example given in

Figure 1, it can be seen how the expression *fraught with* very generally premodifies something bad, especially of three categories, namely, danger, problems and anxiety (but counterexamples are possible, as in line 3). In studies of lexical grammar, concordances are generally used to discover patterns of *collocation*, that is, how any particular word or expression cooccurs with other words or sequences of words with particular frequency. These patterns are often not available to unassisted introspection. The relevant point here is that the patterns of collocation exhibited by a lexical item are very much part of its meaning; 'you shall know a word by the company it keeps' (Firth 1957: 11). Indeed modification in its collocational patterns is a way of both defining and detecting change in meaning of a lexical item.

For studying features of discourse, lists of concordance lines longer than those shown in Figure 1 are generally employed. They may be of several hundred characters (equivalent to text extracts) since a wider context is needed to know what is actually being communicated interactively by the use of the word or expression being examined.

Moving on to the studies themselves, Taylor (2010) examines the changing rhetorical role of *science* in UK broadsheet newspapers from 1993 to 2005. She discovered how the expression *the science* , in formulations such as *the science says / suggests*, *the science shows / reveals*, *the science tells us that* and *as the science demands*, is increasingly used as a model of authority in suasive argumentation. A close reading of the concordance lines revealed however that 'the authority is asserted but relatively rarely justified', that is, it is seldom accompanied by any evidence (2010: 221). Expressions such as (generally unnamed) *scientists revealed / warned / announced*, and so on, were also used more frequently in the 2005 material with the same rhetorical illocution of 'adding authority'. Taylor suggests that this easy appeal to 'the science', especially to banal topics such as *the science of sensible drinking / cooking / shopping* is, one the one hand, a recognition by journalists of the prestigious appeal of science but, on the other hand, likely to lead to a trivialisation of the concept of science. She also notes that the antagonists to *science*, that is, the entities to which UK journalists regard it as being in opposition has altered. In 1993 they were *art* and *culture*, whilst in 2005 they have become *religion* and *ethics*.

Marchi (2010) explores changes in the meaning patterns surrounding whatever is explicitly labelled by the newspapers themselves as some sort of moral issue in the SiBol 93 and SiBol 05 corpora. She begins by ascertaining what news issues are characterised as involving morality and those involving ethics.

She prepared concordances of all items with the stem **moral** (which captured instances in context of *moral*, *morally*, *morality*, *immoral*, and *immorality*) and those with the stem **ethic** (*ethical*, *ethically*, *ethics*, *unethical* and *unethically*), then created a word list for each of the concordances and contrasted them using the keywords tool as described in 1.2 above. This treatment of large concordance files as if they were corpora ('concordance-corpora') and their subsequent subjection to keyword analysis is known as the concordance-keywords technique (Taylor 2010: 226–7).

The *ethic* set of lexical items was found to be preferred when news stories are about business, economy and finance, life-styles (consumption, shopping, fashions and so on), environmentalism and science.

The *moral* set of items is preferred, instead, when news is about religion, feelings and virtues, didactic art (moral *tales*, *stories*, *plays*, *art*, *films*, and so on), people (for example news about *fathers*, *families*, *women*), the political sphere (*war*, *terrorism*, politicians' behaviour) and sexuality. Between 1993 and 2005 Marchi also found a decline in the use of the *moral* set of items (from 99.1 to 67.8 occurrence per million words) but a rise in the use of the *ethic* set (from 28.1 to 41.1 occurrence p.m.w.). Talking about 'morality' is perhaps becoming old-fashioned being replaced by the more fashionable 'ethical'.

To date, there have been two lexical studies which employ three sister corpora. The first is an investigation into the changing discourses on antisemitism in the UK press from 1993 to 2009 (Partington 2012), which draws on data from SiBol 93, SiBol 05 and a specially constructed corpus of all articles from the same three broadsheet newspaper in the year 2009 (downloaded from *Lexis Nexis*). Considerable changes were noted between the discourses in the earlier corpus compared to the later ones. In the first, the majority of these were either historical and/or literary-artistic (typically discussing whether a particular writer or artist had been antisemitic) or, if they were related to contemporary

society, they were discussions of potential or reported antisemitism outside the UK, especially in Eastern Europe. In the later corpora, however, there is much more discussion about a perceived resurgence of antisemitism in the UK and Western Europe.

Partington first prepared a keyword concordance-corpus from each of the three corpora. To identify what they might have in common a keywords comparison of each of the three with an external reference corpus (namely the two-million word British National Corpus sampler, a corpus containing a wide variety of text-types) was conducted. Unsurprisingly, the main common lexis: *hatred*, *Hitler*, *Holocaust*, *Nazi – Nazis*, *racist – racism*, indicated an enduring association of antisemitism with right-wing politics and racism. However, when comparing the later concordance-corpora with that from 1993, other recent discourses surrounding antisemitism emerged to take their place alongside the right-wing associations. There was debate over antisemitism entering mainstream politics with election candidates reportedly making references to the Jewishness of opponents to attract antisemitic votes. There were discussions of reported antisemitism amongst extremist sections of Western Europe's Muslim population and there were discussions of an alleged resurgence of antisemitism on the left. This latter ranged from conspiracy theories of Jews using their clandestine power to control world finance and US policy to reports of harassment of Jewish university groups by left-wing students and of whether some on the left apply an 'affinity for Israel' test to Jewish people (Freedland *Guardian* 2009), when similar tests for other minorities would be considered racist.

This perhaps raises a controversial point. If the nature of antisemitism and the characteristics of its perpetrators have changed, does this necessarily imply that the *meaning* of the lexical item *antisemitism* has altered? Does it not still continue to occupy the semantic space assigned to it by a dictionary description such as "…."? Although this core semantic meaning remains, Partington's study shows that the pragmatics of the word are changing for users, and this is reflected in the lexical patterns in which it is found, and a large part of the meaning of a lexical item is bound up with these ways in which it is used. If the associations of an item, witnessed in the collocations and co-textual patterns, alter significantly over time, then this too is meaning change.

The second study to use three sister corpora, Taylor (2011), is the most explicitly methodological in focus of these modern diachronic lexical studies and uses the three corpora in the SiBol suite (1.2) as the basis of an investigation into ways in which not only changes but also similarities or stasis may be studied in both lexical patterns and socio-cultural attitudes over time. The tools and concepts she examines include:

- Consistency analysis. *WordSmith Tools* allows for the creation of consistency lists when producing word lists, which will identify words which are shared across a number of texts
- Key key-words. 'A "key key-word" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is' (Scott 2008, Users Guide). Thus, items found to be key key-words across texts in the SiBol corpora would indicate consistency of frequency over time.
- Consistent (or wide distribution) collocates. Also known as 'c-collocates', this concept was introduced in Gabrielatos & Baker 2008 to describe the lexical items which collocated, that is, occurred in close proximity in texts, with *refugees / asylum seekers / immigrants / migrants* (RASIM) in at least seven out of the ten annual subcorpora which they had collected from 1996 to 2006. The consistent collocates were calculated in order to exclude collocates which may have been triggered by particular events rather than being representative of newspaper discourse across the extended time period.
- Despite its title the *Sketch Difference* tool, part of the *Sketch Engine* suite of programs,[3] allows the user to analyse similarity as well as difference because it displays the shared collocates as well as the different ones of any two lexical items or expressions the researcher is examining.

As Taylor stresses, there is a natural human tendency to notice change and difference, which is reflected in scientific research. Change can demand an explanation in a way that sameness often does not

3. www.sketchengine.co.uk

seem to and, in conducting research, reporting difference can often appear more exciting than reporting non-change. Indeed much research sets out to look for differences between entities and systems and their behaviours and that is consequently exactly what the analyst is likely to observe and report. And yet such findings are potentially highly misleading since it may be that in both quantitative and qualitative terms, the similarities between two corpora or topics considerably outweigh the differences, and this might well be the most important aspect of the relationship between two or more systems. Taylor (2011) is a reminder to give consistency, sameness, stasis its due regard.

3. Modern diachronic corpus-assisted methodologies: discussion and conclusions

As can be divined from the preceding research outlines, modern diachronic corpus-assisted studies typically employ a combination of both statistical (or 'quantitative') and qualitative techniques, that is, the researcher often moves from overall frequency analyses to close textual reading, often passing through a phase of concordancing, which represents a halfway house in that it involves both statistical-analysis (frequency counting) and text-perusal.

Ellis (1986: 248–76) distinguishes between two distinct (but in practice complementary) modes of conducting research. The first method is induction, or what Ellis glosses as 'the research-then-theory' approach, in which the researcher (i) selects a phenomenon for investigation (we might add, with a research question in mind), (ii) collects a relevant data-set, (iii) looks inside the data-set for systematic patterns and finally (iv) formalises significant patterns as rules describing natural events.

The second mode is hypothesis-driven, what Ellis calls 'the theory-then-research approach', in which the researcher (i) develops an explicit theory (better, hypothesis; and we can add that there is an infinity of ways of arriving at the hypothesis), (ii) derives a testable prediction from the hypothesis, (iii) conducts a research to test the prediction, (iv) modifies (or abandons) the hypothesis if the prediction is disconfirmed and (v) tests a new prediction if the first prediction is confirmed.

Different modern diachronic studies will avail themselves of these modes to differing degrees. Most work into language change is highly inductive, starting from a close analysis of the comparative keywords generated by comparing the lists of items from parallel corpora from different time periods. The researcher is very generally on a journey of discovery, she sets out with little precise idea of what the data will actually tell her. For example, we certainly had no idea which modals, prepositions and linkers were going to prove to be more or less popular over this time period, or even that there was going to be any change in their patterning at all. The observations are very clearly data-driven.

Often the discourse studies, that is, investigations into the changing lexis relating to social and cultural issues of the kind discussed in section 2, are also largely inductive in nature. Another example would be Duguid's (2010b) examination of the *anti* prefix which starts out with the very general research questions of finding out what negative prejudices the SiBol newspapers discuss and whether the targets of the prejudice and opposition expressed by *anti* change over the time period. Again, when setting out, she had little idea what the eventual results would be (for instance, as regards discussions of group prejudices, in 1993 *anti-white*, *anti-European* and *anti-German* proved relatively frequent compared with *anti-American*(*ism*), *anti-British*, *anti-Muslim*, *anti-Catholic* and *anti-Israel*(*i*) in the 2005 data).

Other discourse studies however have more intuitive origins and are driven by somewhat more precise hypotheses, for instance, Taylor 2010 incorporates inductive data-driven statistical analyses of the collocational behaviour in the item *science* and *scientists* but it begins from a hypothesis that 'science, or, more specifically, *the science*, is increasingly being used as a dogmatic model of authority in all spheres of life' (2010: 221).

It is, then, quite natural in corpus-assisted diachronic language research to move back and forth between statistical-quantitative and qualitative analyses and also to combine inductive and hypothesis-driven phases in the same study.

The SiBol suite of corpora has been made publicly available on the *Sketch Engine* website[1] so that other researchers can both conduct their own corpus-assisted modern diachronic investigations and also to check or replicate the work conducted by the SiBol group.

References

- Aarts, Bas, Joanne Close & Sean Wallis. 2010. Recent changes in the use of the progressive construction in English. In Bert Cappelle and Naoaki Wada (eds.), *Distinctions in English grammar, offered to Renaat Declerck*, 148–67, Tokyo: Kaitakusha.
- Baker, Paul. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14:3 312-337.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Bolinger, Dwight. 1975. *Aspects of Language*. New York: Harcourt Brace Jovanovich.
- Carter, Ronald & Mike McCarthy. 2006. *The Cambridge Grammar of English*. Cambridge: CUP.
- Davies, Mark. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+). *International Journal of Corpus Linguistics*, 14:2, 159-90.
- Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus. *Literary and Linguistic Computing*, 25:4, 447-464.
- Duguid, Alison. 2010a. Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora* 5 (2), 109–38.
- Duguid, Alison. 2010b. Investigating *anti* and some reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS). *Corpora* 5 (2), 191–220.
- Ellis, Rod. 1986. *Understanding Second Language Acquisition*. Oxford: Oxford University Press.
- Fairclough, Norman. 1995. *Media Discourse*. London, Arnold.
- Firth, James, L. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.
- Fries, Charles C. 1940. On the Development of the Structural Use of Word-Order in Modern English. *Language* 16, 199–208.
- Gabrielatos, Costas & Paul Baker. 2008. Fleeing Sneaking Flooding. A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics* 36 (1), 5–38.
- Labov, William. 2006 [1996]. *The social stratification of English in New York City*, 2nd edn. Cambridge: Cambridge University Press.
- Leech, Geoffrey. 2004. Recent grammatical change in English: data, description, theory. In Karen Aijmer & Bengt Altenberg (eds.) *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora* (ICAME 23), 61–81. Amsterdam: Rodopi.
- Leech, Geoffrey & Nicholas Smith. 2006. Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. In Antoinette Renouf & Andrew Kehoe (eds.), *The Changing Face of Corpus Linguistics*, 186–204. Amsterdam: Rodopi.
- Lyons, John. 1981. *Language and Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Mair, Christian. 1995. Changing patterns of complementation and concomitant grammaticalisation of the verb help in present-day English. In Bas Aarts and Christian Meyer (eds.), *The Verb in Contemporary English*, 258–72. Cambridge: Cambridge University Press.
- Mair, Christian, Marianne Hundt, Geoffrey Leech & Nicholas Smith. 2002. Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7(2), 245–64.
- Mair, Christian & Geoffrey Leech. 2006. Current Changes in English Syntax. In Bas Aarts and April McMahon (eds.), *The Handbook of English Linguistics*, 318–42. Blackwell: Malden, MA.
- Marchi, Anna. 2010. 'The moral in the story': a diachronic investigation of lexicalised morality in the UK press. *Corpora* 5 (2), 161–89.
- McNair, Brian. 2003. *News and Journalism in the UK*. London: Routledge.
- Morley, John & Alan Partington. 2009. A few Frequently Asked Questions about semantic – or evaluative – prosody. *International Journal of Corpus Linguistics* 14:2, 139–58.
- Partington, Alan. 2010. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora* 5(2), 83–108.
- Partington, Alan. 2012. The changing discourses on anti-semitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study. *Journal of Language and Politics* 11(1), 51–76.
- Scott, Mike. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.
- Taylor, Charlotte. 2010. Science in the news: a diachronic perspective. *Corpora* 5(2), 221–50.
- Taylor, Charlotte. 2011. Searching for similarity: The representation of boy/s and girl/s in the UK press in 1993, 2005 and 2010. Talk given at CL2011, Bimingham University, 22nd July 2011.
- Wells, John. 1982. *Accents of English: An Introduction*. Cambridge: Cambridge University Press.