Language Evolution and the Spread of Ideas on the Web: A Procedure for Identifying Emergent Hybrid Word Family Members

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Liz Price
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: Liz.Price@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321859

Word usage is of interest to linguists for its own sake as well as to social scientists and others seeking to track the spread of ideas, for example in public debates over political decisions. The historical evolution of language can be analysed with the tools of corpus linguistics through evolving corpora and the web. But word usage statistics can only be gathered for known words. In this article, techniques are described and tested for identifying new words from the web, focussing on the case when the words are related to a topic and have a hybrid form with a common sequence of letters. The results highlight the need to employ a combination of search techniques and show the wide potential of hybrid word family investigations in linguistics and social science.

Introduction

There are many situations where it is useful to be able to track the spread or flow of ideas within large groups of people, including advertising campaigns, scientific fields, political debates and newsworthy events. Hence advertisers, politicians, journalists and others need timely information in order to react quickly to events. They probably use a wide range of relatively informal intelligence gathering techniques, from talking to people randomly in the street and reading newspaper letters pages to organised focus groups and opinion polls (e.g., Brookes, Lewis, & Wahl-Jorgensen, 2004; Somin, 2000). Some researchers and others wishing to track ideas use formal text analysis methods (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004; Weare & Lin, 2000). Although hyperlinks have been used a kind of mass plebiscite on the value of individual websites or pages (Brin & Page, 1998; Lifantsev, 2000), text analysis seems to be a more sensitive instrument for tracking ideas because a large number of links is required in order to mine reliable information (Thelwall & Harries, 2004). In fact language use is also of interest in itself as an ideological battleground in addition to its ability to reflect ideas. One high profile example is the "politically correct" phenomenon, much reported in the press. On closer inspection, this debate has been shown to be used as a vehicle for party politics, at least in the UK (Johnson, Culperer, & Suhr, 2003). Moreover, language has also been claimed to be a potential driver of social change, or a means of maintaining the status quo (e.g., Kiesling, 2003; Matsuda, Lawrence, Delgado, & Crenshaw, 1993; Ochs, 1992).

One special case of idea tracking is through the use of individual families of related new words. This can occur when new words are formed as an endemic part of the spread of ideas. In fact, some authors have argued that there is an unnecessary trend towards creating new 'trendy' word families from prefixes (Hale & Scanlon, 1999; McFedries, 2004), as illustrated by the following casual coining of the term 'infoprefixation'.

> "This desire to see things in information's light no doubt drives what we think of as "infoprefixation." Info gives new life to a lot of old words in

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology © copyright 2005 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

> compounds such as infotainment, infomatics, infomating, and infomediary" (Brown & Duguid, 2000, p.3).

There many other examples of new hybrid word families because portmanteau word creation (building new words by combining two old ones) is a standard journalistic and marketing device (van Mulken, 2003); an effective way of succinctly conveying information and contributing to the power of a message through the work the reader has to do to decode it (McQuarrie & Mick, 1996; Meyers-Levy & Malaviya, 1999; van Mulken, 2003). Portmanteau or hybrid word creation is closely related to the simple but powerful advertising technique of juxtaposing two contrasting images that the advertiser wishes the consumer to mentally connect (Phillips & McQuarrie, 2003, 2004; Sonesson, 1996; van Mulken, 2003). We call collections of words created for and centred on an idea *hybrid word families* if they contain a common string of letters that expresses the essence of the idea (e.g., info). In languages such as German in which amalgamated word formation is normal, there will be large groups of hybrid word families, and the families themselves will probably also be large in comparison to English. Given the importance of hybrid word families in many contexts, there is a need to be able to effectively study them.

In order to study hybrid word families, their members must first be found. Systematic identification techniques are therefore desirable, especially for families that arise spontaneously in different contexts. The web forms a natural starting point, and blogs in particular, since it is produced by a large number of people and hence contains a large and quantity of heterogeneous text. Both researchers and businesses (Gill, 2004; Gruhl et al., 2004) have already exploited web text extensively for idea tracking. Nevertheless, existing approaches are restricted to studying individual words or groups of known words related to the ideas being tracked. This will be sufficient in many cases, but not when the word forms themselves are unknown, as in the case of hybrid word families. Direct keyword searches in commercial search engines cannot be used to find all words containing the identifying hybrid word family segment (e.g., *franken*) because they only return whole word matches (or synonyms or plurals). Other techniques, such as word stemming (Porter, 1980) or collocation (Mitkov, 2003) are also likely to be able to identify at least some members of hybrid word families, but both are likely to only capture predictable words (stemming) or words used in predictable circumstances (collocation). Hence there is currently a gap in knowledge: an inability to track the contemporary evolution of hybrid word families.

In this paper we use searching and indexing techniques from information science to construct a multiple method approach for identifying and tracking hybrid word families. A case study of frankenfood words is used to illustrate the approach, and other examples alluded to, suggesting areas in which hybrid word families may be usefully studied. In genetically modified (GM) debates the term *frankenfood*, derived from *Frankenstein* and *food*, incorporates Frankenstein as a metaphor for the dangerous artificial creation of new life by science. This topic is chosen as a real application of the technique, being part of a project to track public science policy debates.

Social aspects of language use

The importance of language for human communication has resulted in many different research areas developing their own language-centred methodologies, addressing a variety of issues. Some of these are briefly reviewed here, forming the background to the methods discussed in the next section and the subsequent case study. Although language is primarily spoken, most of the relevant research reviewed below is of written forms of communication, which have significantly different sets of properties (known as registers) (e.g., Biber, 2003). Web documents will come from a range of styles, with blogs probably tending to be relatively informal, some being close to spoken language. In contrast, academic web sites contain large collections of very formal documents, such as research papers in e-journals, online copies of computer documentation, and

university rules and regulations. Nevertheless, they also contain less formal genres such as personal home pages.

Science communication

The field of science communication is concerned with the communication of science-related information to the public, principally through science journalism. Hence, it is explicitly concerned with the spread of ideas through language. Popular science articles in magazines and newspapers are a common primary research source, in addition to other methods such as surveys of the public (Weigold, 2001). As an example, one study examined the effect of information content of a science news story on readers' perceptions of its believability (Corbett & Durfee, 2004). Another compared newspaper coverage of a large set of articles with their citation rates, finding relationships between the two phenomena (Kiernan, 2003). Content analysis of newspaper and magazine articles is a common research method used in this field.

A bibliometric technique has been harnessed to investigate public science communication, Leydesdorff's co-word analysis. In two recent papers, it was applied to see how word use relating to scientific debates differs across public and academic domains (Hellsten & Leydesdorff, 2005; Leydesdorff & Hellsten, 2005, to appear).

One relevant topic of interest in science communication and science and technology studies is the ways in which scientific ideas can capture the popular imagination in a wider context than justified by their scientific basis: "ideas, models, or theories that have been transposed from their own discipline onto another or from the science to the non-scientific subsystems or everyday discourses" (Maasen & Weingart, 1995, p.17), which probably occurs because of the privileged position of the scientific paradigm in western society (Fuchs, 1992, ch. 1). Here specific words are often a vehicle (or a metaphor) for spreading ideas. Major examples of academic ideas that have caught the public imagination and have influenced the way in which non-scientific activities take place include Darwinianism, chaos theory and Freudianism (Maasen & Weingart, 1995), in addition to those, such as info (meaning computing?) which are dignified by a prefix word family. On a smaller scale, academic ideas frequently get significantly adapted for non-academic environments, one example being the concept of practice, when used in modern business contexts (Vann & Bowker, 2001). Note also that some scientific words, such as catalyst, are brought into general use in a way that probably divorces them almost completely from their original meaning. In such cases word usage may not reflect genuine public recognition of a scientific idea.

Bibliometrics

In bibliometrics, a branch of information science that uses quantitative measures of aspects of document collections, words have been directly used to track ideas within the scientific knowledge production system. Leydesdorff (1989) has shown that the words found in the titles of journal articles can form the raw data for algorithms that cluster articles into sub-fields. This co-word analysis has the interesting property that the words used are automatically extracted by the algorithm used (from article titles) and, unlike in typical corpus linguistics approaches, the words do not need to be explicitly found in advance. The method can be applied to any collection of documents with titles (or equivalent) to reveal its ideational structure (c.f., Leydesdorff, 1997).

Sociolinguistics

Sociolinguistics can perhaps claim to be the archetypal research field for connecting language use to social context, typically in order to explain language use and evolution (Cameron, Frazer, Harvey, Rampton, & Richardson, 1992). Methods are normally qualitative, often including interviews with speakers, and the focus is usually on spoken language (e.g., Hasund &

Stenstrom, 1996). As a branch of linguistics, however, sociolinguistics tends to deal predominantly with language use rather than language as a tool to track the spread of ideas.

The importance of portmanteau words is widely recognised in linguistics, and taught in schools, where the playful possibilities can be exploited (Blachowicz & Fisher, 2003, p234; McKenna, 1978), as they are by some recognised literary figures (Attridge, 1988; Deleuze, 1990, p. xiii). Linguistic discussions of the social implications of hybrid word family formation or portmanteau word construction, which would fall within the realms of sociolinguistics, are rare, however. One exception is an allusion to inter-language portmanteau term formation in Scandinavia, using it as evidence for an inherently multilingual mode of communication (Braunmüller, 2002). Portmanteau words are frequently discussed in linguistics, but mainly from the perspective of small words such as pronouns and prefixes (e.g., *a el* contracted to *al*). Portmanteau word formation is also sometimes used to describe the construction of words from meaningful lexical units smaller than words (i.e., morphemes, see http://en.wikipedia.org/wiki/Morpheme), and are closely related to clitics (words that have no independent function apart from other words, such as *las* in *las aguas*, see http://en.wikipedia.org/wiki/Clitic).

Corpus linguistics

Language and its underlying social factors can be studied by taking a large collection of text and applying various qualitative and quantitative techniques to draw conclusions. This is the realm of corpus linguistics. For instance, the association of swearing in English with lower social classes has been claimed to be an outcome of the seventeenth century bourgeois revolution (McEnery, 2005). McEnery uses a corpus linguistics approach: extracting word frequency statistics from relevant bodies of text together with contextual information in order to construct his argument. Corpus linguistics is often contrasted to Chomskian linguistics, which relies upon human intuition and understanding of language rather than statistical analysis: both have their place in contemporary language study (McEnery & Wilson, 2001). Although a range of standard corpora, such as the British National Corpus (BNC), are used to analyse many different aspect of language use, they tend to be static, remaining unchanged since their completion date (BNC, 2004). Some, such as the COBUILD corpus used to help build language use dictionaries, do evolve continuously but are not able be compiled and published fast enough to keep up to date with the most recent evolving language trends. As a result, the web has come into use as an enormous de-facto linguistic corpus, using commercial search engines as its interface (Davies, 2001; Mair, 2003; Meyer, Grabowski, Han, Mantzouranis, & Moses, 2003).

A limitation of search engines for linguistic analysis is that they only allow whole word searches, and linguists are frequently interested in semantically related words (via a process called lemmatisation, rather than hybrid word families). In response, a search engine interface has been developed to automatically submit a series of related queries to search engines to include a range word combinations desired by a linguist, presenting the results in a form suitable for corpus linguistics (Fletcher, in press). This is not sufficient to *identify* members of hybrid word families, however, since it relies upon predicting word forms and submitting queries for them to see if they can be found.

A method for hybrid word family usage identification

The hybrid word family identification task is defined to be the discovery of words that conform to a given hybrid word family pattern, as defined above. This section describes a range of different methods for identifying hybrid word family members.

Word searches by predicted forms

For search engines that do not allow searches within words, such as for word stems, it may be possible for a researcher to predict likely portmanteau words based upon known examples of usage and their origins (etymology). For example, words related to a pop star might be expected to be derived from an opinion (e.g., Britneyphile, Britneyphobia, Britneylover, Britneyhater) or from previous experience with similar word formations (e.g., familiarity with the word beatlemania may suggest searching for Britneymania to see if the same word formation pattern has been used). Based upon this premise, words can be predicted and their actual existence tested for on the web through standard searches. This is similar to Mair's (2003) technique for assessing the existence of rare grammatical formations that are predicted by a theory-driven hypothesis to exist.

This method should be effective at finding predictable word constructions, as long as the number of possible word combinations to be tested is not too large, but cannot find unexpected word combinations.

Wildcards in search engines

Some search engines allow keyword searches to include wildcards to represent missing parts of words. These used to be offered by many search engines including AltaVista, Inktomi (iWon), Northern Light, Yahoo! and AOL Search (Sullivan, 2003, March 11) but none of the major search engines offered wildcard searches as of November, 2004. The recent trend towards concentrating the core search capability in a small number of search engines, and the increasing cost of starting new search engines (Van Couvering, 2004) have combined to create a system where the exact form of search allowed is dependant upon a small number of search engines. Experiments with all of the search engines listed by SearchEngineWatch.com revealed only one that still offered a wildcard search, dmoz.org. In its search guide (http://dmoz.org/searchguide.html), it explains that this can be used only at the end of words so that `Bicycl*` "would match sites on Bicycling, Bicycle, and Bicycles", but, "The search does not support arbitrary wildcards, so searches on "*cycling" or "Arch*ology" will not work." Nevertheless, this is a powerful command to find new hybrid word forms. Dmoz's 4-million site coverage (as of November, 2004) is impressive, but its search coverage is typically restricted to a single page per site. Dmoz, the open directory project, is a non-profit initiative that does not crawl the web but uses human indexers to categorise sites.

Note that WebCorp (http://www.webcorp.org.uk/) also offers a wildcard search facility via commercial search engines, but its results are extracted from non-wildcard searches. For example, a search for Franken* could be translated into a Google search for Franken but in the results from Google all matches for Franken* would be highlighted.

Authority identification

Statistical analyses of web patterns for many web-related phenomena, including page sizes, link counts and word usage patterns often show small numbers of documents dominating, an aspect of the 'power law' effect (Baldi, Frasconi, & Smyth, 2003; Barabási, 2002; Levene & Poulovassilis, 2004; Thelwall, 2005, to appear; Zipf, 1949). This suggests that an effective strategy for finding hybrid word family members would be to look for individual pages that contain many such terms, which may be seen as authorities on the topic and hence likely to contain many different related words. Identifying and scanning authority pages would therefore be a productive way to find new word forms. It is common sense that in many cases the use of one word family member in a document makes it more likely that other family members will also be mentioned in the same document. Following links or otherwise identifying related documents is also a logical way to look for uses of similar words, an information seeking process sometimes called chaining (Ellis, 1989). This method may help to find unpredicted word uses that are

nevertheless related but is unlikely to produce examples of words formed in completely different contexts. This latter information may be particularly sought as evidence of a deeper cultural embedding of the concept associated with the word family.

Vocabulary or index searching

Every search engine must maintain a vocabulary, a list of all terms extracted from all of its web pages. The construction of a vocabulary is an essential part of indexing, the process that makes keyword searching possible (Chakrabarti, 2003). If a commercial search engine published its index, then this could be easily searched for occurrences of relevant hybrid words (e.g., by a simple text search of the vocabulary). Unfortunately, however, this perfect solution to the hybrid word family problem is not (yet) possible because commercial search engine vocabularies are not published.

An alternative is to use a personal web crawler and personal search engine. These can be used to crawl and index a set of web sites, producing a vocabulary for those sites. The main limitations of this approach are that it takes time and resources, and that it would be impractical to crawl many web sites because it would take too long. Nevertheless, this strategy may be capable of producing examples of hybrid word construction in unexpected contexts, if enough web sites are crawled. A web crawler, complete with basic tools for vocabulary creation and searching, is available at http://socscibot.wlv.ac.uk/.

Although web sites represent a very wide range of types of publication, characterised as "the loose web" (Burnett & Marshall, 2002), they are probably still generally representative of a more formal style of language use. Web logs (blogs) probably tend to incorporate language more towards the informal side of the spectrum (Nardi, Schiano, Gumbrecht, & Swartz, 2004) and with younger authors (e.g., Huffaker & Calvert, 2005; Kumar, Novak, Raghavan, & Tomkins, 2004), even though the most popular examples tend to be created by a highly literate, educated section of the population (Gill, 2004; Matheson, 2004). Blogs are based upon a technology that allows casual web users to easily publish online, in the form of a continuously updateable series of dated log entries. These are used for various purposes, from keeping a very public personal diary to attempting to create an authoritative resource for a given topic (Bar-Ilan, 2004; Matheson, 2004). Personal blogs are particularly interesting from a sociological perspective, being perhaps the largest-scale publicly accessible source of personal opinions (Sunstein, 2004). Blogs, therefore, are probably one of the most accessible large-scale sources of data about the spread of ideas in society (Glance, Hurst, & Tomokiyo, 2004; Schiano, Nardi, Gumbrecht, & Swartz, 2004). Although harvesting the contents of a large number of blog postings is technically challenging because of the variety of web page formats used, their key content can often be accessed in a simplified XML form via the RSS (Really Simple Syndication) feed technology (Glance et al., 2004; Gruhl et al., 2004). Essentially, this allows researchers to identify a group of blogs and then use RSS reader software to periodically monitor and report new blog postings. Vocabularies constructed from RSS feeds are therefore an especially useful source of new words.

Summary

All of the methods described above have limitations, either in their coverage of the web, or in their power to extract words from the web. Hence, since their weaknesses only partially overlap, the best strategy is to employ all of them and combine their results.

Online word usage tracking methods

This section describes methods for using the web to track the evolution and use of groups of words over time. These techniques are not specific to hybrid word families.

Word usage estimates

Any known word can be searched for in a commercial search engine like Google. The count of the number of results pages (e.g., "Google found about 1,230 matches") gives an approximate indication of the frequency of use of the term. This is essentially the same approach as used in previous corpus linguistics studies (Blair, Urland, & Ma, 2002; Meyer et al., 2003), and suffers from the same limitations such as the potential for the counts to be influenced by the creation of large numbers of pages by single authors. It is important to note that the statistical word frequency results returned by search engines are not reliable or always accurate because they return counts of the number of pages containing one or more occurrences of a word, and these counts may include multiple copies of similar pages, or examples of the word used in different pages, but in a repeated section of the page (e.g., a standard links bar, or company address line) (Fletcher, in press). Moreover, web text represents web publishing and non-web documents subsequently posted to the web, presumably tending to exclude other forms of language, such as spoken conversations.

Word usage tracking

The evolution of word usage over time can be tracked with search engines that allow date-specific searches. For example, at the time of writing, AltaVista allowed searches to be restricted to pages last modified by any given date range (Leydesdorff & Curran, 2000). This is particularly useful to identify early examples of new word usage but note that the date refers to the last modified date of the page, rather than its creation date, so other pages may be missed if they are not indexed by the search engine or have been subsequently modified. A potentially better source of temporal web information is the Internet Archive, which records and saves web pages permanently, providing a historical record of the web (Burner, 1997). At the time of writing its keyword search interface looked highly promising but was not yet fully working.

Geographic usage spread

Search engine data can provide some evidence of geographic spread, e.g. Australia vs. the U.K (Ingwersen, 1998). A simple way to estimate national differences in the use of a word is to search for the word twice, each time requesting the search engine to return only pages from one country, a service that many search engines offer, either directly or indirectly through top level domain searches (e.g., for pages with domain names ending in .uk). Considerations such as the total volume of web publishing in different countries should be taken into account, however, as should the possibility that the search engine is incorrectly assigning a significant number of pages to the wrong country, which is especially likely in the case of top-level domains (Smith, 1999). Hence large differences in relative frequencies of use would be needed for reasonable evidence of a genuine difference, and this would need to be supported by manual checking of results, as discussed below.

Results validation

In order to provide some reliability assurance of the web data used for the above three purposes, a random sample of pages in each reported category should be visited to check that they are what they purport to be (Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998; Thelwall, 2004). The original total can be multiplied by the proportion of the random sample that is correct in order to get a corrected estimate of the number of pages containing the word. For word usage frequency reporting, counting duplicate copies of a page should be excluded but these are hard to identify through the random sampling of matching pages because both original and duplicate would have to fall within the random sample. An alternative method would be to take a more investigative (and less scientific) route, searching for likely sources of duplicates in the results. This should be used if it is suspected that very few matches are genuine, otherwise the statistical

estimates may be very poor. To give a simple example, if a word occurs in 100,000 pages but only 100 times in the correct context then a random sample of 10 from the 100,000 would probably give no correct matches, hence 'predicting' 0 matches in the entire 100,000.

In the geographic results an additional source of anomalies is the 'incorrect' assignment of country codes to pages. This can be remedied by a straightforward random sample of each country code TLD and multiplying the overall result by the percentage of correct TLD assignments. For example if 1000 pages were found from the French .fr top level domain, but in a random sample of 50, 90% were found to come from France then the figure should be modified to 1000 x 90/1000 =900 pages. The remainder can be reallocated to the correct domain.

Case study: Frankenscience words

In 1999 a public debate into the use of genetically modified (GM) food emerged when governments discussed the introduction of GM crops, with protestors raising concerns. Language has been an issue almost from the moment newspaper started to cover the story. As part of the debate, the US Food and Drug Administration unsuccessfully tried to get the press to use the terminology genetically *engineered* food (McInerney, Bird, & Nucci, 2004, p. 70), seeing the word *modified* as problematic. As part of the protest strategy, the GM foods were dubbed 'frankenfoods', merging the words Frankenstein and food. This portmanteau word was intended to suggest the bad consequences possible from scientists modifying and reconstructing life forms (a classic image juxtaposition device). The frankenfood metaphor was popularised in the press and gave rise to a Franken- hybrid word family, including related portmanteau words like frankenscience and frankencrops. A good illustration of the use of hybrid words to politicise language is the following extract from the Greenpeace web site "Genetically Engineered fish threaten world's oceans - Greenpeace calls for GE-free seas. TAKE ACTION: Say No To Frankenfish!" (http://archive.greenpeace.org/geneng/highlights/gmo/GEfish.htm). Tracking public use of franken family words can therefore be used as a method to investigate an important aspect of the GM food debate. Any GM-related word starting with franken is casting GM food in a negative sense, and so its use would be presumably welcomed by the protesters and discouraged by GM supporters.

The GM food debate has also been analysed in the science literature because of its strength in the sense of being able to influence governmental science policy (Clarke, 2003; Klintman, 2002; Levidow, 2001; Yearley, 2001), in contrast to the more common situation of "politics [being represented] as a sport played by political and media elites" (Brookes et al., 2004). This upset has lead to the trialling of new forms of democratic participation in science policy debates (Hagendijk, 2004), and much of the literature has alluded to the frankenfood metaphor (Tait, 2001; Ten Eyck & Williment, 2003).

The web seems a natural choice for studying the GM debate (McInerney & Bird, 2005). Frankenscience words have previously been studied on the web (Hellsten, 2003) through the following AltaVista search (AltaVista's wildcard is now discontinued).

```
frankenfood* OR (frankenstein food*)
```

Hellsten (2003) found that the frankenfood metaphor was highly used in various related web sites in 1999 (e.g., Friends of the Earth and the Times), but tailed off after this. In the web as a whole, however, the number of relevant pages increased exponentially, and did not tail off after 1999. It is unclear whether this indicates a widening debate, or a continuation of the debate outside of its prime site. The debate was still live in 2004-5 through activists (e.g., www.greenpeace.org) science debate web sites (e.g., at www.publicdebate.com.au) and GM-specific information sites (e.g., www.genewatch.org, www.gmsciencedebate.org.uk).

Word identification

The following steps were taken to identify franken word family members from the web.

1. *Predicted word forms* New frankenscience words were predicted by forming new words from those already known. This lead to the construction of possible new words from the names of cereal crops.
2. *Authority page identification* A search for two of the known frankenscience words was used to find pages with a high frequency of frankenscience word use. An authority page URL was found, a log of news stories posted by the Organic Consumers Association (http://organicconsumers.com/log.html), with brief summaries of each one. A browser search using Internet explorer's internal find feature with the text "Franken" was also used.
3. *Wildcard searching* Dmoz was searched with the query franken* and the results pages scanned to identify new frankenscience word forms.
4. *Vocabulary searches 1* Web crawls were conducted of the authority site identified in step two (http://organicconsumers.com/), using a publicly available web crawler and search engine (http://socscibot.wlv.ac.uk/) described in detail elsewhere (Thelwall, 2001, 2004). The vocabularies created were loaded into the Windows Notepad text editor and searched, using Notepad's Find function, for the text string "franken".
5. *Vocabulary searches 2* U.K. and Australian university web sites were chosen as examples collections of web sites not directly related to the GM issue. The choice of academic sites was a practical one: first, these sites had already been crawled and so the data could be reused; second, there are difficulties with obtaining permission to crawl large non-academic web sites. Crawls of both sets of sites in 2003 and 2004 produced a combined word list with a total of 93 words containing "franken".
6. *Vocabulary searches 3* A collection of 3,702 Blog feeds were harvested from 25 Nov to 17 Dec 2005, producing a total of 109,501 individual postings with 6,661,019 individual terms. Two sources were used to build this collection: the syndic8.com site, which lists a large number of feeds, and Google Advanced searches `filetype:rss` which search specifically for the RSS file name extension. Note that neither source is perfect: the former contains self-submitted feeds, and the latter ignores feeds with filenames ending in .xml and .rdf, which appear to form a significant proportion. Hence the collection is ad-hoc.

During the search, it became clear that there were four different types of word: GM-related words; non-GM words alluding to amalgamation; Frankenstein words; and Germanic words. All words found were investigated using Google searches and categorised into one or more of the above groups, based upon their contexts of use in the whole web (not just where they were originally found). The results of the first two types are summarised below in Tables 1 and 2. The non-GM foods in Table 2 exhibit a range of variation in degree of Frankenstein-relatedness, from frankenpotato (a potato decorated to look like Frankenstein) to frankenputer, with franken meaning amalgamated, probably in a self-depreciating sense. Frankenstein words were all direct derivations such as frankensteinian. Germanic words were related to places in Germany and family names of German origin. Compound word formation is common in German and Germanic words were categorised as German rather than listed in Table 2. An example of this was Frankenpost, a newspaper. All of the words which contained franken after the start were Germanic (mainfranken, meierfrankenfeld, mittelfranken, oberfranken, unterfranken) and found in the academic corpora.

The wildcard search produced a word that was not in actual use but only in the URL of the page: http://members.home.nl/frankenkarin/happy/ was the web site of "Frank en Karin" (Dutch language). This was ignored, as were spelling mistakes. Plural forms of words were manually converted to singular when both occurred.

Table 1. GM-related words containing "franken".

Word	Code	Context
frankenfood	1,2,3,4,5au,6	GM debate term
frankencorn, frankenwheat, frankencrop	1,2,4	GM food type
frankenscience	1,5au	GM debate term
frankenfish, frankensoya	2,4	GM food type
frankengrass, frankenseed, frankentree	2,4	GM non-food
frankenrice	2	GM food type
frankenfries	3	GM food type
frankenbuck, frankenfight, frankengene, frankenclone	4	GM debate terms
frankentony	4	GM processed food mascot
Frankenbean, frankenfruit, frankengrape, frankenlettuce, frankenpig, frankensalmon, frankenspud, frankenwine	4	GM food type
frankencotton, frankendrug, frankenfarm, frankenfleece, frankenlawn, frankenpharm, frankenplant, frankenweed, frankenkitten	4	GM non-food
frankenburger	5au	GM food (also a family name)
frankenfear	5u	GM debate term (fear of GM food)
frankenpet, frankenrunner	5u	GM non-food

*1=predicted, 2=authority (organicconsumers.org log page), 3=wildcard, 4=organicconsumers.org crawl, 5=crawls (a=Australian universities, u = UK universities), 6=Blog feeds

Table 1 shows very clearly that the identified authority site was a major source of frankenfood words. This fits the profile of an information war, in which organicconsumers.org is promoting the use of frankenscience words as part of its protest strategy. The site contains lists of press releases, many of which discuss new GM issues. Interestingly, despite its distinctively authoritative status, there were a significant number of GM words that it had not used, or were no longer on its site. The relationship between the site and the origins of the stories is not always clear: for example 'frankentony' seems to have been invented by Greenpeace as a device to highlight the potential use of frankencrops in breakfast cereals (Greenpeace, 2004).

The difference between the typical origins of words in tables 1 and 2 is clear: GM words come from an activist site, whereas non-GM words tend to originate in the academic sites. This does not implicate academics in the creation of the latter words, the dominance of the academic web sites over blog feeds, for example, is almost certainly mainly due to their enormously greater size. From Table 2 it is also clear that the Frankenstein metaphor is by no means restricted to the GM debate, and that there is a playful theme running through the words. This is stronger than the gothic theme that would naturally be associated with the Frankenstein story. There is also a minor computing theme, which probably reflects the large number of computing pages in academic web sites more than the gothic sense of humour of some computer scientists.

Table 2. Words containing 'franken' but unrelated to GM issues.

Word	Code*	Context
frankenfish	2,4	A term used to describe a monster-like fish.
frankenhooker	3	Film (inspired by the Frankenstein story).
frankenshred	3	Rock band name.
frankenstudent	3	Cartoon student.
frankenbok	5a	Australian rock group.
frankenfido	5a	Frankenstein joke.
frankenmazda	5a	One person's car rebuilding project.
frankenskippy	5a	Cartoonist's pseudonym.
frankentoon	5a	Frankentoon (combination of "Frankenstein" and "cartoon").
frankenbundle	5au	A system for maintaining and distributing text formatting software.
frankenchrist	5au	An album released by an alterative rock group.
frankencamera	5u	A hybrid camera, built from different sources.
frankenclone	5u	A hybrid computer, built from different sources.
frankenham	5u	Pig joke: "The bride of Frankenham" (also family name check)
frankenjack	5u	A cocktail - word origin unknown.
frankenpotato	5u	Potato dressed up to resemble Frankenstein.
frankenputer	5u	A hybrid computer, built from different sources.
frankenbike	6	A hybrid bike, built from different sources.

*1=predicted, 2=authority, 3=wildcard, 4=organicconsumers.org, 5=crawls (a=Australian universities, u=UK universities), 6=Blog feeds

Word usage

The known franken words were searched for in Google, and the results shown in figures 1 and 2, using a logarithmic scale because of the expected power law. The results in Figure 1 are in most cases the raw Google page counts, with only a few modifications from the results of the random samples. The random sample exercise was problematic because many of the pages came from blogs where they were parts of postings repeated in follow up blog comments. These were left in because it was difficult to identify this phenomenon precisely and a blog posting follow-up represents a display of interest in the blog, and probably the franken posting. The following changes were made: frankenburger reduced from 4,420 to 2 (most were family names; about 10 were references to a picture of a Frankenstein statue next to a fast food restaurant, without any GM references); frankenbean reduced from 117 to 2 (most were family names, others mentioned a children's play "Frankenbean and the Monster Carrots"): frankenfish reduced from 66,000 to 5,000 (most pages referenced monster fish types). The following changes were made to figure 2: frankenbean reduced from 117 to 60: frankenfish reduced from 66,000 to 61,000; frankenham reduced from 29 to 2. The word frankenfish was used significantly in GM and non-GM contexts, and is in both graphs.

The straight line graph in Figure 2 reflects a power law that is common in word frequency statistics (Li, 1992; Zipf, 1949) and found almost everywhere on the web (Baldi et al., 2003; Barabási, 2002; Thelwall, 2005, to appear), but the non-linear Figure 1 is more surprising, showing that there is an unnaturally high number of low frequency GM words. This would be consistent with a set of activists attempting to promote the usage of certain words, but these words not achieving a resonance with the public. This could be thought of as an unnatural phenomenon from the perspective of language evolution. It seems reasonable to conclude from the graph that the only unambiguously GM related word that has achieved a significant usage rate is frankenfood. This is clearly a value judgement, but even frankenscience seems to be used about ten times less than the name of an album by a little-known old pop group (frankenchrist) and a B-movie (frankenhooker).

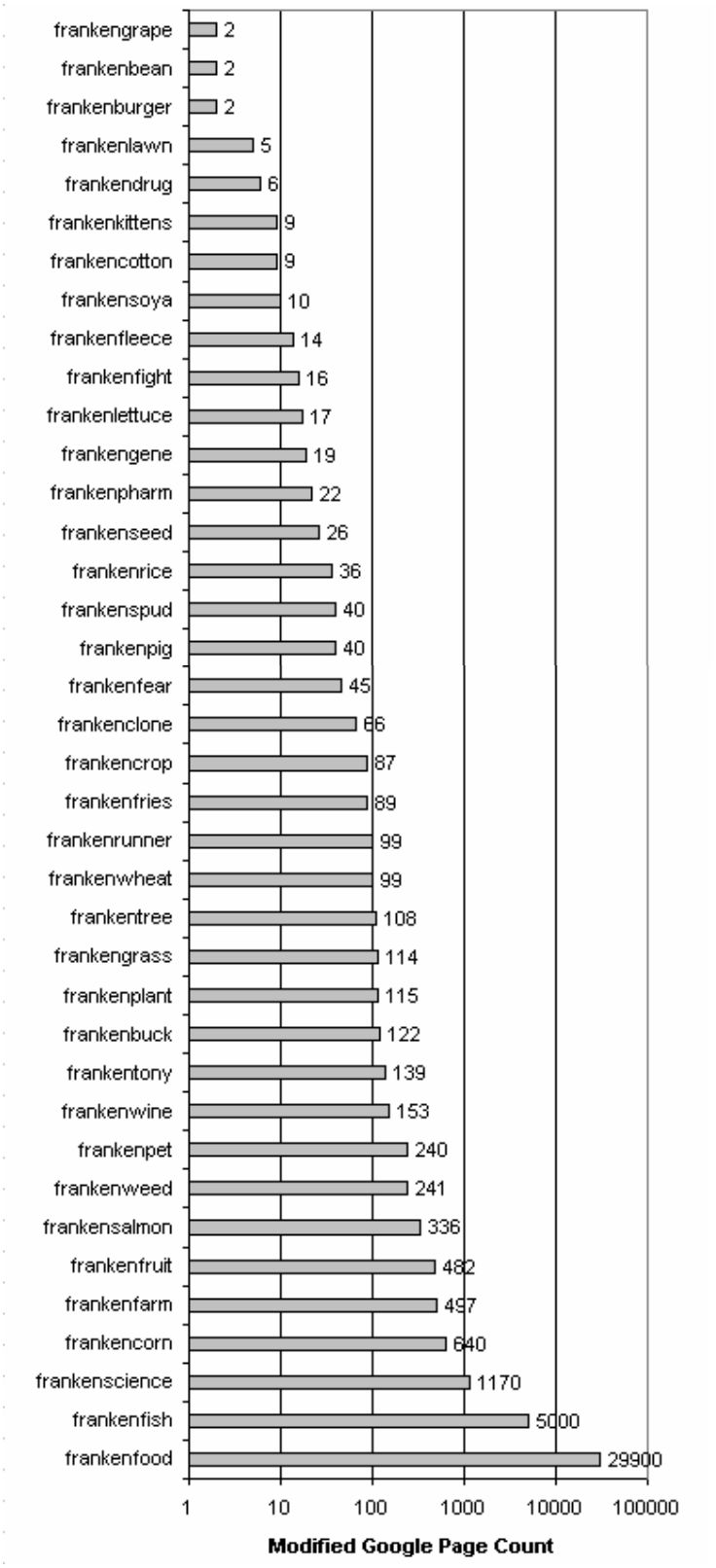

FIG. 1. Pages in Google containing GM-related words with the stem franken.

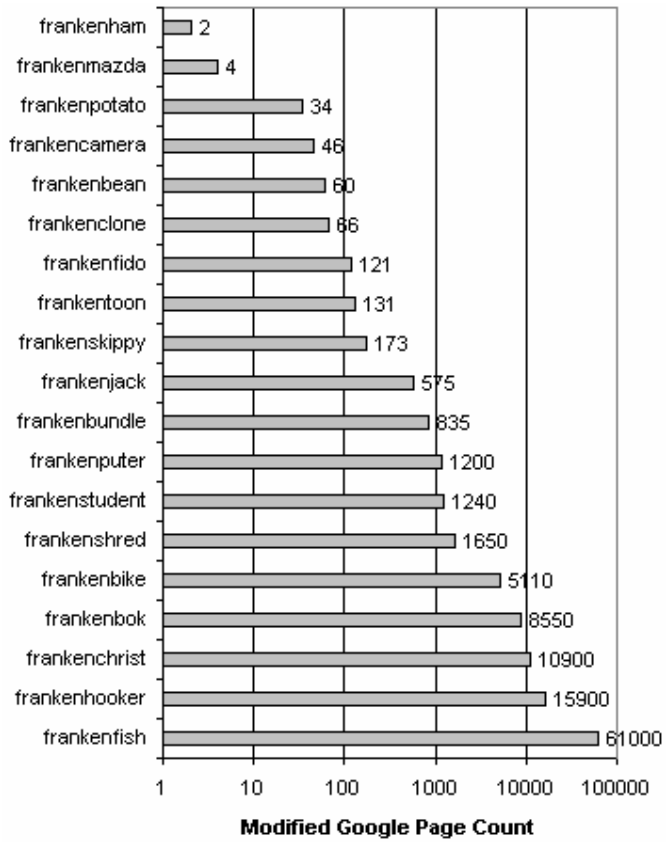

FIG. 2. Pages in Google containing non-GM words with the stem franken.

Geographic spread

Figure 3 shows the top-level domains of pages containing the term frankenfood, illustrating the kind of information that can be revealed. The graph is difficult to interpret because .com and .org sites could have any origin. The random sampling of the country-specific domains gave no examples of incorrectly placed pages, although the international domain pages would probably have a natural country of location in most cases. Interestingly, in the non-English speaking countries, the term frankenfood was found predominantly in native language pages, often in quotes. This shows that the importance of the term itself to the debate, as well as its international nature.

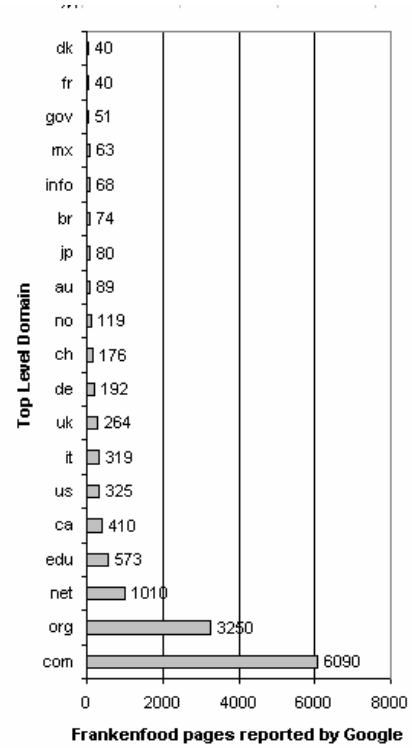

FIG. 3. The top 20 TLD sources of pages containing the term Frankenfood, as reported by Google in December 2004.

Other case studies

The statistics reported above were also calculated for an eclectic collection of other hybrid word families (not shown). The results of some are summarised very briefly here, providing clues as to when the techniques of this paper will give useful results.

- **Web families** Hybrid word families containing the internet-related terms *blog*, *web*, *cyber* and *net* were investigated. The principle difference with frankenscience was that there was an enormous collection of words in each case because of their web-specific nature and the number of web sites that had deliberately constructed hybrid names in order to have a name that was also a domain name (spaces are not allowed in domain names). A linguistically interesting phenomenon was that the cyber- word family was predominantly related to law and crime, whereas such meanings were rare in the other categories. It seems that 'cyber' has strong connotations with illegal activity.
- **Science families** Hybrid families containing *nano*, *bio* and *astro* were chosen as common scientific root words. In fact, their commonness was a problem as in the web-related word family case, generating long lists of words. Other, more specialised terms, such as *bioinf* yielded a more manageable set of results. For example, the UK corpus contained 52 bioinf family members (e.g., macrobioinformatic).
- **Word endings** A search for words ending in *ism* found a very large numbers of terms, 5,566 in the UK corpus alone, but produced a list of words that may have been of interest to sociologists.
- **Pop culture** Searches for Beatle, Britney and spice families found some in each case, including beatlemania and beatlemaniac, although less than expected. This probably reflects a lack of relevant selected sources: some music-related sites could have improved the results. Amongst the other pop culture searches tried, searches for queer family words

gave the most varied results, including queercore, queerable and shakesqueer. This seems to indicate a conscious attempt to play with language.

Discussion of limitations

Hybrid word family identification will be useful only in specific cases, like those above, where families are naturally created. Although there are many examples of hybrid word families, noun clusters are probably far more common, since they can easily be built by adding new words to an existing concept. An example in science is "computer assisted assessment".

Hybrid word family identification

The main case study validates the method for finding hybrid word family members in the sense that the six different methods all identified at least one word that the other methods did not (with the exception of method 1, of course), as shown in Table 1. Note that method 4 should naturally subsume method 2, but did not because the crawler was not able to extract all the words from the web pages crawled. The case study has limitations in generalisability, highlighted by the fact that some of the other examples produced word families that were either too large to tackle with the same techniques or too small to be worth investigating. In a sense it is an almost ideal case, with franken being relatively rare in English words, Frankenstein being a loan word from Germany, where Franken- is a more common word stem. The other case studies showed that very small and very large hybrid word families also exist. Nevertheless, the strength of this case study is that it is a genuine application of the techniques, to aid a non-linguistic purpose, tracing public engagement in a science policy debate. A more subtle limitation of the study is that the use of a web crawler raises ethical concerns since it consumes the resources of the web site crawled (web server time and bandwidth (Koster, 1993)). Partly to combat this, keyword lists from large academic crawls have been placed free online (http://cybermetrics.wlv.ac.uk/database/). Nevertheless, the methods will be most effective when a key source site can be identified and crawled.

Word spread

The word usage and geographic spread graphs (figures 1-3) gave revealing information about the word family, including the hypothesis that GM-related words were exhibiting an artificial frequency distribution, identifying geographic centres of the debate and the multilingual use of the term frankenfood. As discussed above, page counts reported in Google are not reliable indicators of word usage; they must be interpreted cautiously because they can be greatly affected by spurious factors. Nevertheless, it seems that it would be difficult to get better estimates because constructing a robust language usage corpus is a time-consuming process (Burnard, 1995).

One limitation of the approach described is temporal: although Google discards old pages when they disappear or are replaced with new ones, the page count statistics reported could include text created from any time when the web existed, and in some cases contains archives of older texts converted to web form. As a result of this, the date on which the pages were created is uncertain and it is not possible to gain high quality time series statistics. In response to this issue, we are currently compiling a large blog corpus that will contain accurately time-stamped data and give an unprecedented ability to make accurate estimations of word usage changes over time.

Conclusions

A method has been described to identify and track the evolution of hybrid word families, using the web. The case study has shown that it is both practical and capable of giving useful information. The combination of different techniques was useful, with commercial search

engines offering wide web coverage and the personal web crawler giving the opportunity to find new words created out of topical context. The limitations of the approaches are reported above, but the most important is that the only type of language that can be investigated is that of web publishing: incorporating many varieties of written texts, but mostly excluding speech and being unrepresentative of written text in general. Nevertheless, these techniques give an opportunity to systematically study evolving language use for hybrid word families in a way that has not been possible before.

The success of the method points to its potential use outside of a science policy context. From the literature review, other relevant topics would be portmanteau words in advertising (perhaps even developing techniques to identify the rise of new trendy prefixes and to predict their demise), science mapping, general journalism, other languages (e.g., German) and linguistics (e.g. basic word construction from morphemes). In each of these areas the availability of relatively straightforward methods to identify hybrid word families makes their study possible for the first time.

Acknowledgements

The referees are warmly thanked for their helpful suggestions. The work was supported by a grant from the Sixth Framework for Research and Technological Development of the European Commission. It is part of the CREEN project (Critical Events in Evolving Networks).

References

Attridge, D. (1988). Unpacking the portmanteau, or who's afraid of Finnegans Wake. In J. Culler (Ed.), *On Puns: The Foundation of Letters* (pp. 140-155). Oxford: Basil Blackwell.

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modelling the Internet and the Web*. Wiley: Chichester, UK.

Barabási, A. L. (2002). *Linked: The new science of networks*. Cambridge, Massachusetts: Perseus Publishing.

Bar-Ilan, J. (2004). An outsider's view on "topic-oriented" Blogging. *World Wide Web Conference*, http://www.www2004.org/proceedings/docs/2002p2028.pdf.

Biber, D. (2003). Variation among University spoken and written registers: A new multi-dimensional analysis. In P. Leistyna & C. F. Meyer (Eds.), *Corpus Analysis: Language Structure and Language Use* (pp. 47-70). Amsterdam: Rodopi.

Blachowicz, C. L. Z., & Fisher, P. (2003). Keep the "fun" in fundamental: Encouraging word awareness and incidental word learning in the classroom through word play. In J. F. Baumann & E. J. Kameenui (Eds.), *Vocabulary Instruction: Research to Practice* (pp. 218-239). Guilford: Guilford Publications, Inc.

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34*(2), 286-290.

BNC. (2004). What is the BNC? , http://www.natcorp.ox.ac.uk/what/index.html.

Braunmüller, K. (2002). Semicommunication and accommodation: Observations from the linguistic situation in Scandinavia. *International Journal of Applied Linguistics, 12*(1), 1-23.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Brookes, R., Lewis, J., & Wahl-Jorgensen, K. (2004). The media representation of public opinion: British television news coverage of the 2001 general election. *Media culture and society, 26*(1), 63-80.

Brown, J. S., & Duguid, P. (2000). *The social life of information*. Boston: Harvard Business School Press.

Burnard, L. (1995). *Users' reference guide to the British National Corpus*. Oxford: Oxford University Computing Services.

Burner, M. (1997). Crawling towards eternity: Building an archive of the World Wide Web. *Web Techniques, 2*, http://www.newarchitectmag.com/archives/1997/1905/burner/.

Burnett, R., & Marshall, P. (2002). *Web theory: An introduction*. London: Routledge.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. London: Routledge.

Chakrabarti, S. (2003). *Mining the Web: Analysis of hypertext and semi structured data* . New York: Morgan Kaufmann.

Clarke, B. (2003). Report: Farmers and scientists: A case study in facilitating communication. *Science Communication, 25*(2), 198 - 203.

Corbett, J. B., & Durfee, J. L. (2004). Testing public (un)certainty of science. *Science Communication, 26*(2), 129-151.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science, 49*(14), 1319-1328.

Davies, M. (2001). Creating and using multi-million word corpora from web-based newspapers. In R. C. Simpson & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp. 58-75). Ann Arbor: University of Michigan.

Deleuze, G. (1990). *The logic of sense*. New York: Columbia University Press.

Ellis, D. (1989). A behavioral approach to information retrieval systems design. *Journal of Documentation, 45*(1), 171-212.

Fletcher, W. (in press). Making the web a more useful source for corpus linguistics. In U. Connor & T. Upton (Eds.), *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. Amsterdam: Rodopi.

Fuchs, S. (1992). *The professional quest for truth: A social theory of science and knowledge.* Albany, NY: SUNY Press.

Gill, K. E. (2004). *How can we measure the influence of the blogosphere?* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). *BlogPulse: Automated trend discovery for weblogs.* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Greenpeace. (2004). Kellogg's creates a monster. http://web.archive.org/web/20040111062110/http://www.greenpeaceusa.org/features/monster.htm.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through Blogspace.* Paper presented at the WWW2004, New York, http://www.www2004.org/proceedings/docs/1p491.pdf.

Hagendijk, R. (2004). Framing GM food: Public participation and liberal democracy. *EASST Review, 23*(1), 3-7.

Hale, C., & Scanlon, J. (1999). *Wired Style: Principles of English Usage in the Digital Age*. New York: Broadway books.

Hasund, I., & Stenstrom, A.-B. (1996). *Conflict talk: A comparison of the verbal disputes between adolescent females in two corpora.* Paper presented at the ICAME 17, Stockholm.

Hellsten, I. (2003). Focus on metaphors: The case of "Frankenfood" on the web. *Journal of Computer Mediated Communication, 8*(4), http://www.ascusc.org/jcmc/vol8/issue4/hellsten.html.

Hellsten, I., & Leydesdorff, L. (2005). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Frankenfoods,' and 'stem cells.' *in preparation*, http://users.fmg.uva.nl/lleydesdorff/meaning/measuring%20meaning.pdf.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2), http://jcmc.indiana.edu/vol10/issue12/huffaker.html.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation, 54*(2), 236-243.

Johnson, S., Culperer, J., & Suhr, S. (2003). From 'politically correct councillors' to 'Blairite nonsense': discourses of 'political correctness' in three British newspapers. *Discourse & Society, 14*(1), 29-47.

Kiernan, V. (2003). Diffusion of news about research. *Science Communication, 25*(1), 3-13.

Kiesling, S. (2003). Prestige, cultural models, norms & gender. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender*. Oxford: Backwell.

Klintman, M. (2002). The genetically modified (GM) food labelling controversy: Ideological and epistemic crossovers. *Social Studies of Science, 32*(1), 71-91.

Koster, M. (1993). Guidelines for robot writers. Available at: http://www.robotstxt.org/wc/guidelines.html.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.

Levene, M., & Poulovassilis, A. (Eds.). (2004). *Web Dynamics.* Berlin: Springer.

Levidow, L. (2001). Precautionary uncertainty: Regulating GM crops in Europe. *Social Studies of Science, 31*(6), 842-874.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*, 209-223.

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science, 48*(5), 418-427.

Leydesdorff, L., & Curran, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. *Cybermetrics, 4*, http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html.

Leydesdorff, L., & Hellsten, I. (2005). Metaphors and diaphors in science communication: Mapping the case of 'stem-cell research', *Science Communication*, 27(1), 64-99.

Li, W. (1992). Random texts exhibit Zipf's-Law-like word frequency distribution. *IEEE Transactions on Information Theory, 38*(6), 1842-1845.

Lifantsev, M. (2000). Voting model for ranking Web pages. In P. Graham & M. Maheswaran (Eds.), *Proceedings of the International Conference on Internet Computing* (pp. 143-148). Las Vegas: CSREA Press.

Maasen, S., & Weingart, P. (1995). Metaphors - Messengers of meaning. *Science Communication, 17*(1), 9-31.

Mair, C. (2003). *Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora.* Paper presented at the ICAME conference, Guernsey.

Matheson, D. (2004). Weblogs and the epistemology of the news: Some trends in online journalism. *New Media & Society, 6*(4), 443-468.

Matsuda, M., Lawrence, C., Delgado, R., & Crenshaw, K. (Eds.). (1993). *Words that wound: Critical race theory*. San Francisco: Westview Press.

McEnery, A. M. (2005). *Swearing in English*. London: Routledge.

McEnery, A. M., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McFedries, P. (2004). The (pre) fix is in. *IEEE Spectrum Online,* http://www.spectrum.ieee.org/WEBONLY/resource/aug04/0804tech.html.

McInerney, C., & Bird, N. (2005). Assessing Website quality in context: Retrieving information about genetically modified food on the Web. *Information Research, 10*(2), http://InformationR.net/ir/10-12/paper213.html.

McInerney, C., Bird, N., & Nucci, N. (2004). The flow of scientific knowledge from lab to the lay public: The case of genetically modified food. *Science Communication, 26*(1), 44-74.

McKenna, M. C. (1978). Portmanteau words in reading instruction. *Language Arts, 55,* 315-317.

McQuarrie, E. F., & Mick, D. G. (1996). Figures of rhetoric in advertising language. *Journal of Consumer Research, 22*(4), 424-461.

Meyer, C., Grabowski, R., Han, H.-Y., Mantzouranis, K., & Moses, S. (2003). The world wide web as linguistic corpus. *Language and Computers, 46*(1), 241-254.

Meyers-Levy, J., & Malaviya, P. (1999). Consumers' processing of persuasive advertising: An integrative framework of persuasive theories. *Journal of Marketing, 63*(45-60).

Mitkov, R. (2003). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

Ochs, E. (1992). Indexing gender. In A. Duranti & C. Goodwin (Eds.), *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge, UK.: Cambridge University Press.

Phillips, B. J., & McQuarrie, E. F. (2003). The development, change, and transformation of rhetorical style in magazine advertisements 1954–1999. *Journal of Advertising, 31*(4), 1-13.

Phillips, B. J., & McQuarrie, E. F. (2004). Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing Theory, 4*(1-2), 113-136.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Schiano, D. J., Nardi, B. A., Gumbrecht, M., & Swartz, L. (2004). *Blogging by the rest of us.* Paper presented at the Conference on Human Factors and Computing Systems (CHI 2004), Vienna.

Smith, A. G. (1999). A tale of two web spaces; comparing sites using Web Impact Factors. *Journal of Documentation, 55*(5), 577-592.

Somin, I. (2000). Do politicians pander? *Critical Review, 14*(2-3), 147-155.

Sonesson, G. (1996). An essay concerning images: From rhetoric to semiotics by way of ecological physics. *Semiotica, 109*(1/2), 41-140.

Sullivan, D. (2003, March 11). Search Features Chart. *SearchEngineWatch*, http://searchenginewatch.com/facts/article.php/2155981.

Sunstein, C. R. (2004). Democracy and filtering. *Communications of the ACM, 47*(12), 57-59.

Tait, J. (2001). More Faust than Frankenstein: The European debate about risk regulation for genetically modified crops. *Journal of Risk Research, 4*(2), 175-189.

Ten Eyck, T. A., & Williment, M. (2003). The national media and things genetic: Coverage in the New York Times (1971–2001) and the Washington Post (1977-2001). *Science Communication, 25*(2), 129-152.

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science, 27*(5), 319-325.

Thelwall, M. (2004). *Link analysis: An information science approach*. San Diego: Academic Press.

Thelwall, M. (2005). Text characteristics of English language university web sites. *Journal of the American Society for Information Science and Technology,* 56(6), 609-619.

Thelwall, M., & Harries, G. (2004). Do better scholars' Web publications have significantly higher online impact? *Journal of American Society for Information Science and Technology, 55*(2), 149-159.

Van Couvering, E. (2004). *New media? The political economy of Internet search engines.* Paper presented at the Annual Conference of the International Association of Media & Communications Researchers, Porto Alegre, Brazil.

van Mulken, M. (2003). Analyzing rhetorical devices in print advertisements. *Document Design, 4*(2), 114-128.

Vann, K., & Bowker, G. (2001). Instrumentalizing the truth of practice. *Social Epistemology, 15*(3), 247-262.

Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web-Opportunities and challenges. *Social Science Computer Review, 18*(3), 272-292.

Weigold, M. F. (2001). Communicating science: A review of the literature. *Science Communication, 23*(2), 164-193.

Yearley, S. (2001). Mapping and interpreting societal responses to genetically modified food and plants. *Social Studies of Science, 31*(1), 151-160.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Cambridge, MA: Addison-Wesley.