A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary

Ruth O'Donovan
Mary O'Neill
Chambers Harrap

For each new edition of The Chambers Dictionary, around 1,000 new words are selected by Chambers' lexicographers for inclusion. In preparing the latest edition, we seized the opportunity to use new corpus and database technology to improve neologism detection and selection. Our resources included the large, recently built Chambers Harrap International Corpus (CHIC), our automated word-tracking system, the databases developed for our new words monitoring programmes and a new tool for ranking words by corpus frequency. We report on the results of our work in this area: a systematic approach to neologism detection and investigation that complements the expertise of lexicographers.

1. Introduction

The task of selecting new words for inclusion in a new edition of a large monolingual dictionary such as *The Chambers Dictionary* is a challenging one. This challenge is twofold. Firstly, any new additions to the language must be observed and recorded. Secondly, prior to inclusion in the dictionary each word must be evaluated on the basis of its longevity and actual usage. It is neither possible nor desirable to include every new word which has come into existence since the previous edition. Neologism selections cannot be made at random, necessitating a systematic approach whereby empirical evidence is used to inform the judgements of experienced lexicographers.

Following collaboration between lexicographical and computational staff we have developed an approach to this task with two related strands. At the heart of both is the extensive, ever growing Chambers Harrap International Corpus (CHIC), which is described in Section 2. In Section 3, we describe a corpus-based automatic word-monitoring program and web interface for the identification and evaluation of neologisms. In Section 4, we introduce our databases of neologisms recorded by lexicographers and other contributors. From these we generate frequency lists (Section 5) to support the lexicographers in their decision-making (Section 6). In Section 7, we discuss the usefulness of the approach and report on lexicographers' observations. In Section 8, we outline the application of this approach to the selection of neologisms for online dictionaries. In Section 9, we conclude and outline plans for future improvements based directly on user feedback.

2. Chambers Harrap International Corpus

Chambers Harrap International Corpus (CHIC) consists of more than 500 million words of modern (post-2002), international English. It is a regularly updated monitor corpus in the tradition of the Bank of English (*http://www.collins.co.uk/books.aspx?group=153*) rather than a static, balanced resource such as the British National Corpus (*http://www.natcorp.ox.ac.uk/*).

Corpus content is compiled in two ways. Every day our web-spidering system collects data from specified online resources such as newspapers, magazines and websites. The downloaded files are HTML and must be cleaned in an initial processing step to convert them to text. HTML tags, JavaScript blocks, image links, menus and other markup is removed automatically. The files are then written to a MySQL database along with details about genre, domain, author and place of publication. Prior to writing a data chunk to the database we run a duplicates check to make sure it has not already been added to the corpus.

In addition to data sourced from the web, we regularly receive published books and newspapers in electronic formats such as PDF, Quark and ACSII from co-operative publishers. If necessary, these files are automatically cleaned and converted to plain text format before being written to the CHIC database with their associated metadata. The metadata labels used for domain, genre and place of publication are listed in Table 1.

Three times a year an updated version of CHIC is built from the entire database to upload to Sketch Engine (Kilgarriff et al. 2004), our corpus query tool of choice. At this stage, the corpus is part of speech (POS) tagged using TreeTagger (Schmid 1994) to allow for more advanced corpus querying. In addition, the system allows the exploitation of metadata labels for the building of specialized subcorpora and filtered concordance searches. The build of CHIC used for the work described here is almost 400 million words in size and was uploaded in June 2007. At that stage, almost 50% of the corpus was made up of newspaper and magazine data, with the UK and the US as the dominant places of publication.

Genre	Place of Publication	Domain
Books: Fiction	Australia	Applied Science and Technology
Books: Non-Fiction	Canada	Arts and Entertainment
Magazine	India	Commerce and Finance
Newspaper	Ireland	Education/Pedagogy
Spoken	Israel	Free Time and Leisure
Website	Malaysia	Generalities
Blog	South Africa	Mathematics and Natural Sciences
	UK	Religion/Spirituality/Philosophy
	US	Social Science
	Mixed	Sport

Table 1: CHIC Metadata labels

3. Automated word tracking

As described in Section 2, our corpus-building efforts involve the regular (daily or weekly) access and retrieval of data from multiple sources. This provides us with a diachronic overview of the language from those sources, in particular facilitating the observation of the emergence and life cycle of neologisms and coinages. To exploit this, the lexicographers and computational team collaborated to build a web-based tool for the regular identification of neologisms for potential inclusion in the dictionary.

Following established approaches to neologism discovery (such as, amongst others, the work of Eiken et al. 2006), we first built a stable reference corpus containing data from a specified set of magazines, newspapers and websites over a period of twelve months. The reference corpus is used as model of "normal" language use. This corpus is POS tagged and lemmatized. At present the POS tags are only used to exclude proper nouns in the absence of a Named Entity Recognizer (NER). The system works at the lemma level, so lemma type frequency lists are generated from the reference corpus.

At the end of every month, we compile data downloaded that month from the same set of sources used to build the reference corpus. This data is then cleaned, POS tagged and lemmatized and used to create the monthly or test corpus. Lemma type frequency statistics are also generated for the monthly corpus. The log likelihood statistic (see for instance Dunning 1993 and Rayson and Garside 2000) is used to identify and rank lemmas occurring with a significantly greater frequency in the monthly than in the reference corpus. Absolute frequency thresholds are used to disregard potentially common lemmas in a first step to optimize the output. A threshold is also set on the log likelihood value of the lemmas. In addition, we wish to exclude words currently in our dictionary, so a headword list augmented with inflected forms acts

as a second filter before the words are written to the database. The inclusion of inflected forms in the filter provides a safety net for potential failures in lemmatization.

The next step is for lexicographers to select words of interest from the automatically generated list, and each month they have in the region of 600 proposed neologisms to examine. To facilitate this, we have developed a web front-end to the database allowing users to interactively assess the output of the neologism tracker (see Figure 1 below). For each new-word suggestion in the list, the user can click through to a KWIC-type citation with further options to link through to concordances in CHIC and ukWaC[1] and hits in Google and Wikipedia. Figure 2 below shows the citation for the suggested new term *hedgie* (hedge fund manager). The links to Google, Wikipedia and the corpora are shown below the citation. The lexicographer uses the coloured buttons to assign a status to the potential neologism: "definitely keep", "monitor further" or "discard"[2]. When he/she has reviewed all the output for a particular month it may be archived using the archive button on the main page (Figure 1), but the user can continue to view and change the status of archived content. All discarded entries are added to an additional filtering stage so that they will not reappear in subsequent months.

Based on numbers alone, the productivity of the automated word tracking system tends to be greater than that of human monitors, simply because it is unlikely to miss a previously unseen word. Thus, it provides a "safety net" against human oversight. However, the output will inevitably be noisy. The noise is sometimes caused by human error in the original data production and so includes, for example, common misspellings such as *teh* for *the* and *recive* for *receive*. Limitations of the current system are also a cause of noisy output. At present our preprocessing stage does not include named entity recognition, something which we hope to incorporate in the future. We depend solely on the tagger's ability to recognize proper nouns. Consequently, results frequently include the likes of *springsteen* and *Harrods*. Occasionally the HTML clean-up stage introduces errors such as the deletion of spaces between words or incomplete tag stripping, resulting in the output of invalid concatenations such as *itemsKate* as seen in Figure 1 below.

Finally, the system will always identify unassimilated foreign words as neologisms. Examples include *maudit*, listed in Figure 1 below. It comes from the following statement: "…he decided to bring Curtis's life to the screen and concentrate on the man rather than the *poète maudit* of popular culture". *The Chambers Dictionary* traditionally excludes unassimilated foreign words that may occasionally appear in English-language material. This is discussed further in Section 4.

Due to these shortcomings of the automated word tracking system, the scrutiny of a lexicographer is required to convert it to a truly valuable resource. The system itself also benefits from this scrutiny as the lexicographers' judgements are fed back into the filtering phase to improve future performance.

[1] ukWaC is a large web corpus available as part of the Sketch Engine package.

[2] The "discard" option is the default as we have found that the vast majority of suggestions will be rejected.

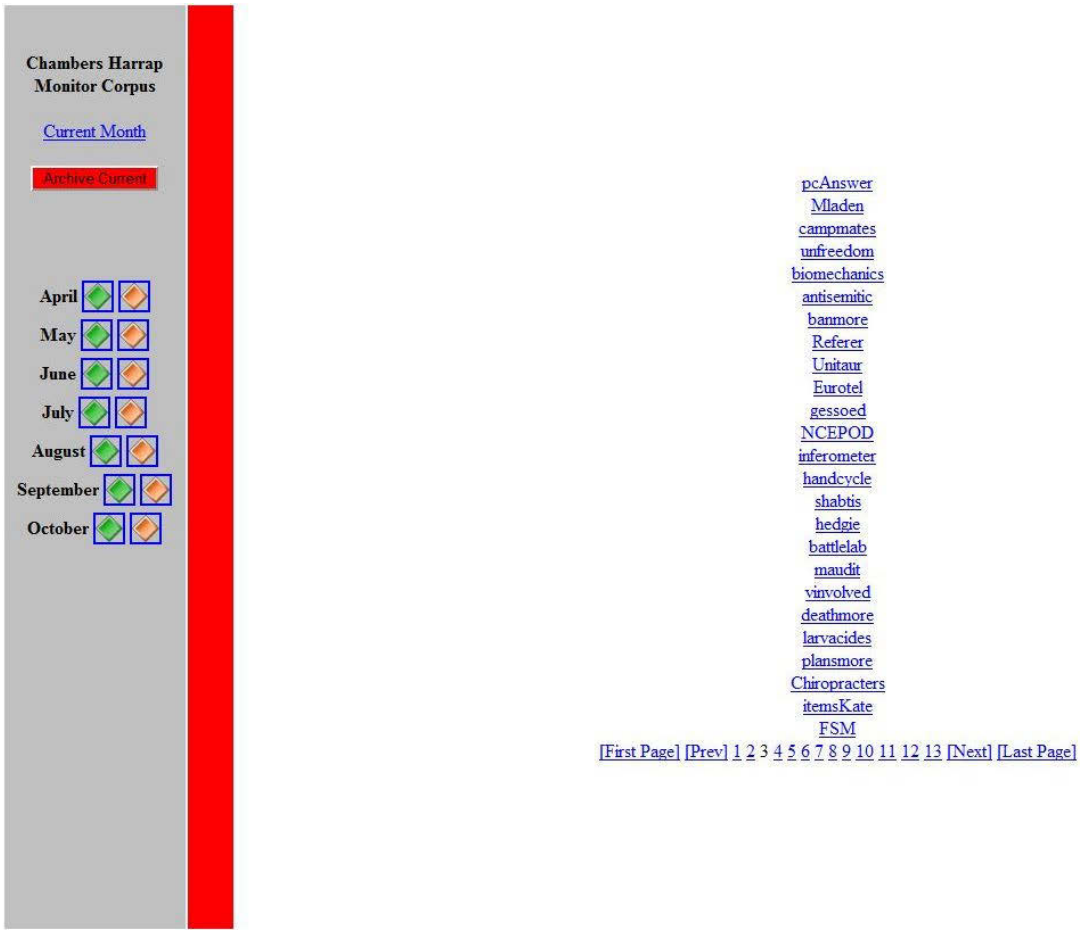

Figure 1: Automated Word Tracking Home Page

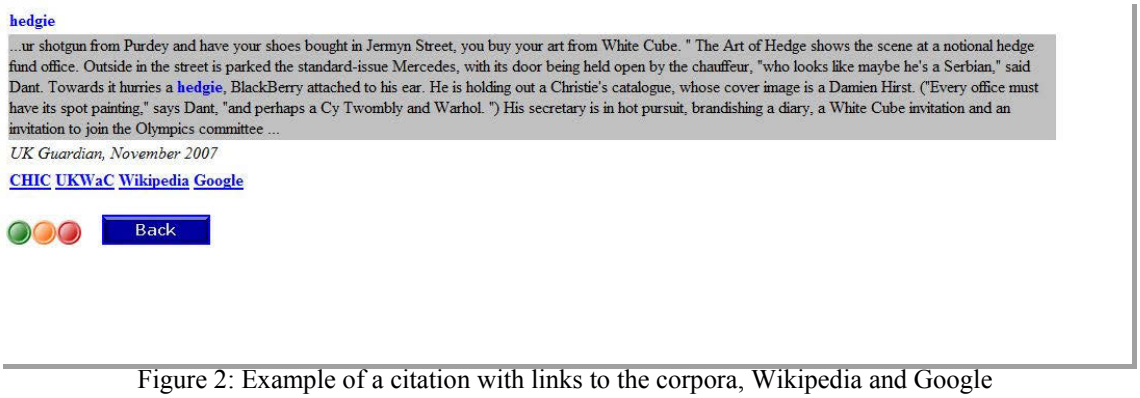

Figure 2: Example of a citation with links to the corpora, Wikipedia and Google

4. Manual word tracking

Wordtrack is a directed reading programme in which commissioned readers scan publications that have been selected as potentially productive sources of new words. The readers' primary task is to identify new words and new senses of words, and every one sent to us is logged on a database, along with source and date information. Approximately 200 new-word citations are recorded each month. Also available to Chambers lexicographers is the Neologisms database, an in-house resource in which editors record several new words and senses every week.

Inevitably, the word lists from these databases have an advantage over the automated word tracking output in that they are relatively junk-free. The technical noise identified in Section 3 above is absent. Also less likely to appear are words that are already included in *The Chambers Dictionary* but with, for example, a different spelling, in a hyphenated form, or without capitalization. For example, the solid compound *roadmap* (with the sense "a plan for achieving

a goal via a series of intermediate stages") appears high on the automated concordance list with 398 citations, but a cursory check reveals that the word was entered in a previous edition of the dictionary, as *road map*. However, the appearance of these variant forms cannot be regarded merely as a distracting flaw in the data: a high number of citations of a word in a different form may lead the lexicographer to conclude that that form is just as, if not more, common than the one included, and he/she can modify the entry accordingly.

These databases are less likely to contain lexical items that fall outside the selection criteria of *The Chambers Dictionary*. As indicated in Section 3, unassimilated foreign words that may occasionally appear in English-language contexts (for example, *ahora, polpette*) will be identified as neologisms by the automated tracking system, but not by dedicated readers or lexicographers who will determine immediately that these are not required in a monolingual English dictionary. However, a case can be made for the recording of such terms: if attempts were made to exclude entirely from the system words from languages other than English, lexicographers could be denied evidence of a foreign word becoming used widely in English-language sources. This is particularly apparent in the field of food and drink: as the cuisine of a country is enhanced by culinary imports, the language is enriched by lexical imports which, though well established in other languages, are gaining new currency in English-language contexts. New arrivals in *The Chambers Dictionary* such as *dashi*, *pad thai* and *vongole*, though arguably not fully assimilated into English, appear with such regularity in English-language menus and food packaging that to omit them from the dictionary might deny users essential information. Viewed in this light, the inclusion of words from other languages might be regarded as an advantage rather than a drawback. The identification and scrutiny of such citations certainly raises interesting issues, largely outside the topic and scope of this paper, of when a foreign word can be considered an integral part of English.

The reader or lexicographer will also be aware that, traditionally, *The Chambers Dictionary* does not include names of people, places or organizations (*Al-Qaida*, *Millennium Dome*); compounds and phrases that are deducible (*clothes swap, a whole lot*); and quotations and phrases that are essentially political slogans (*axis of evil*). Keeping these particular principles in mind, the human reader will not record such items for consideration; the automated process, of course, cannot bring any such judgement to bear on each item. For these reasons, a database of manually recorded words will, inevitably, be more precise and clean than the automated collection, although it is very unlikely to be as comprehensive.

Another considerable advantage of manual over automatic collection of citations is the recording of new senses of an existing word. The new edition of *The Chambers Dictionary* will include for the first time, for example, more recently emerged meanings of the words *blowback* ("repercussions or consequences of an action"), *hot spot* ("an area where a computer user can make a wireless connection to the Internet"), *rip* ("to copy—digital data—from a CD or DVD onto a hard disk") and *spin out* ("of an educational establishment: to create—a company—to exploit commercial opportunities identified by its research") among others. Invariably, new senses are identified in the manual collection programmes. To produce manageable output that is of any practical use, the automated tracking system described in Section 3 has to exclude word forms already included as references in the dictionary, making it a tool that is used solely for the observation of neologisms, and not for new uses of established words.

This is also true of words displaying a shift in the part of speech in which they are used, usually (and to the chagrin of some) from noun to verb. Most often, such uses will emerge in the manual word collections, but these can also be discovered if new inflected forms are picked up in the automated word tracking. An example of such a form as identified by this system is *bullet*, highlighted in a report on a football match: "Kean *bulleted* a header just over". The lexicographer is alerted to a verb usage he/she may have assumed was already covered, but that has not been filtered out against the list of headwords and inflected forms already in *The Chambers Dictionary*.

5. Frequency list generation

Owing to the size of the new words databases, it would be extremely time-consuming for a lexicographer to examine their contents and enter each as a query into the Sketch Engine manually before getting to the important job of sifting through empirical evidence of usage to inform their judgements on the value of the word. We devised a means of automating the first two steps, thus allowing the lexicographers to focus on the task of interpreting corpus evidence. We developed an application that automatically extracts the suggested new terms from each database and constructs a query string for each. This is sent to the Sketch Engine concordance query page. The HTML response is parsed to extract the number of hits for the term in the CHIC corpus. The results are output as HTML files (one for each letter) containing a frequency-ordered list of terms, their actual frequencies and a link to their individual concordance set. If the number of hits in CHIC is five or fewer, a second query is constructed from the term and sent to the Sketch Engine to ascertain the number of hits in the ukWaC resource. The results are output in the same way. These interactive lists are then passed to the lexicographers, who can get an overall view of corpus frequency before simply clicking on a link to further investigate actual instances of use.

6. Lexicographical applications of frequency lists

When put to lexicographical use, the frequency lists alone assist to some extent in the selection of new words. Previously, the lexicographer would have as a starting point a list of neologisms that was as likely to be sorted by alphabetical order as by frequency in an extensive corpus. Frequency ordering presents a more functional view of the candidate list, where items not supported by strong corpus evidence are not given unwarranted prominence. For example, in an output file for letter R, *reimagine* ("to present a fundamentally new interpretation of a subject, especially an artistic work") appears higher on the concordance list than *ringxiety* (the anxiety felt on hearing a mobile ringtone the same as yours when the call is not for you). *Ringxiety* may be an interesting coinage, but with only four corpus citations it is of very little practical value in the world of dictionary compilation. This more balanced view of the data allowed us to impose some guideline frequency criteria for the lexicographers to use in their research. For words that were not technical, abbreviations or proper names, editors were instructed to give serious consideration to words with at least 25 corpus citations; words with between 10 and 25 concordances were found to be less strong candidates for selection, although these were not to be automatically dismissed, for reasons we discuss in Section 7 below.

7. Discussion

Although the basic assumption could be made that words that score highly on corpus frequency will be the strongest candidates for inclusion in a dictionary, it was evident that some of the top-ranking words did not merit inclusion, and that some interpretation of the output would be required in order to locate the lexicographic gold among the items.

Therefore, of even greater value to the lexicographer than raw frequency lists alone are the links from each word to its full concordance set. It is accepted that corpus analysis is a requisite of contemporary lexicography, and it is unthinkable that a word might now be included in a dictionary without corpus research to establish its meaning and the scale of its usage. The links automated a process that otherwise could be time-consuming and frustrating. Each lexicographer analysing a word list was able to disregard "non-dictionary" items as identified in Section 4, and check the concordances for all those higher-frequency words that are identified as potential neologisms. As they did so, they were able to identify points to be borne in mind when assessing such frequency lists for the purpose of selecting neologisms, and raise issues that might be addressed in future refinements of programs linking neologisms lists with corpus frequency.

For example, one of the principles for adding new words to *The Chambers Dictionary*, as for many other dictionaries, is that a word should be in use over a relatively broad area. The lexicographer will be alive to the fact that, even if the absolute frequency of a word suggests it

is worth researching, a single source may account for a large proportion of its occurrences in a corpus. Using the link to corpus citations is essential, as a glance at the concordances page(s) will show, even without further investigation of the metadata, the predominance of a single document-id code. Of the fourteen hits for *cosmeceutical*, for example, four came from an article in *Newsweek*, and ten from one book on the topic of health and beauty. If further evidence is required for the widespread appearance of a higher-frequency word, the Sketch Engine allows a "Sample" option that quickly creates a random sample with a specified number of citations. In the course of our neologism research, this frequently proved a reliable indicator of the range of sources in which a word appeared.

Furthermore, the lexicographer will also be aware that it is not uncommon to find the same text occurring in several different source documents within a corpus, but with each source repeating the content of, for example, a single news story. A brief analysis of the content of the citations establishes if this is the case. (An unfortunate event gaining widespread news coverage in Britain in 2007 was the disappearance in Portugal of a little girl from the UK. A Portuguese term denoting "suspect" status—*arguido*—was used in many reports, often without translation, assuming the reader or listener would have gained awareness of its meaning. However, there were no citations for the word outside the context of this particular story).

It is for a similar reason that scientific terms are treated with care, even when these are supported by high frequency figures in the output. Technical terms are often found widely in the news media after the announcement of a discovery or incident, before disappearing back into the realms of academia. We are inclined to impose a higher than usual burden of proof before selecting names of pharmaceutical products or medical procedures for inclusion, so it becomes crucial to examine the metadata behind the concordances in order to establish the chronological spread of the citations. A frequency list might give a misleading rank to ephemeral items, and these can only be identified from analysis of the citations behind the raw figure. Also thrown up in the corpus frequency lists, perhaps predictably, are many abbreviations. Because of the vast number in existence, abbreviations may be subject to an even higher standard of frequency before being given serious consideration. Moreover, even when an abbreviation does meet raw frequency requirements, further investigation may reveal a spread of several meanings, no single one of which merits inclusion: *RLF*, for example, can be interpreted as be "The Royal Literary Fund", "Revolving Loan Fund", "relaxin-like factor" and more.

Predictably, the raw frequency of a word is of little help when the candidate word is a new sense of an existing word. For example, new senses of *cheese* and *churn* had been identified as candidates for addition, but these words occur so frequently in their standard senses in the corpora that the frequency figures for them are worthless in the selection process. Viewing the concordances, it is difficult to locate those citations where the word has the required sense. Also among our candidate words were several proper names (for example, *Joe Strummer*, *Anna Kournikova*[3]) that have acquired slang usages. Unsurprisingly, these names appeared fairly high in our corpus frequency lists, but their connected citations were seldom in the sense with which we were concerned.

Neither are frequency figures of any value where words are included as dictionary entries to complete a lexical set, regardless of their frequency. For example, if the genetics term *haplotype* is selected for inclusion, the lexicographer would be inclined to add the related term *diplotype*, even if the latter could not be justified purely on the grounds of frequency. Furthermore, there will always be examples of words that may lack overwhelming corpus evidence but that compilers will instinctively recognize as meriting inclusion in a dictionary (the word *fatoosh*, for instance, was familiar to the lexicographers, but surprisingly low on corpus citations). No matter how large a corpus is, there may be gaps in its coverage, and a word appearing low

[3] In this context *Joe Strummer* is rhyming slang for a "bummer", a depressing experience. An *Anna Kournikova* turns out to be a poker hand consisting of an ace and a king, partly from the initials *AK*, and partly because it "looks good but rarely wins anything".

ranking in a frequency list can give a false impression of how common a word is. Ultimately, corpus frequency alone cannot be taken as a hard-and-fast indicator of a word's use; the lexicographer must use his/her judgement in analysing the lists.

8. The new tools and online dictionary provision

The latest edition of *The Chambers Dictionary* will be a central resource in *Chambers Reference Online*, which will be launched simultaneously with the book. In planning and creating the online product, we looked to incorporate the most desirable functional features of the best online dictionaries. In addition, we were aware that users of online dictionaries have high expectations of up-to-date content, and that the traditional three- to five-year update cycle of the printed book would not suffice for a product with the potential to include very recent new words and senses. Therefore, to meet the required level of both currency and comprehensiveness, plans are in place for quarterly updates of the content. Without the automation of the selection processes to facilitate the addition of new material, such regular updates may have been too time- and labour-intensive to be feasible. In addition, the monthly output from the automated tracking system allows lexicographers to extract, and consider including in the online dictionary, words that are truly at the cutting edge of vocabulary.

An online dictionary also has the benefits afforded by the removal of the space constraints of the print product, the most obvious being that more words can be included. Fewer items need be discarded from each edition, or excluded from the outset, on the grounds that they occupy space that would be merited more by another item. The identification of more potential neologisms by the new automated systems—roughly 600 per month from the automated tracking system alone—means that this potential for greater comprehensiveness can be fulfilled. Words which may have been judged as being not sufficiently established or potentially ephemeral can be included. (Words falling into this category in selecting neologisms for the 2008 edition included *telenovela* and *filk*—a science-fiction music genre—). Should any of these words be introduced to the online dictionary and subsequently prove to have the short lifespan predicted, they can easily be removed if the editor chooses, and not become the quaint anomalies that one sometimes finds in print dictionaries.

9. Conclusion

Overall, this systematic approach to neologism selection has been of immense benefit in identifying potential neologisms for dictionary inclusion, and in doing so with greater speed and precision. Automatic frequency rankings, in conjunction with linked concordance lists, provide evidence of a word's currency and usage that can be interpreted by lexicographers to inform their selection of neologisms for dictionary inclusion.

The collaboration between lexicographers and corpus developers is ongoing, and we have plans to improve our processes. For example, in future we hope to include information about corpus distribution across the various metadata categories. We will also investigate the extraction of frequency lists from subcorpora reflecting the balance of subject fields in *The Chambers Dictionary*. We would also like to improve the performance of the automated word tracking system by incorporating a named entity recognizer. At present the system only works purely on a lemma level but we also hope to investigate the possibility of identifying new phrases and compounds. These and other further refinements will give lexicographers much-improved tools to assist with the decisions inherent in dictionary making.

Acknowledgements

The authors are grateful to Ian Brookes for his contribution to the work described here, and to Vicky Aldus and Katie Brooks for their feedback and suggestions.

References

Dictionaries

The Chambers Dictionary. [11th edition.] Edinburgh: Chambers Harrap Publishers, 2008.

Other works

Dunning, T**.** (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19 (1). 61-74.

Eiken, U. C. et al. (2006). "Ord i Dag: Mining Norwegian Daily Newswire". In Salakoski, T. et al. (eds.) Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006 Proceedings. Lecture Notes in Computer Science 4139. Springer. 512-523.

Kilgarriff, A. et al. (2004). "The Sketch Engine". In *Proceedings of EURALEX 2004*. Lorient, France. 105-115.

Rayson, P.; Garside, R. (2000). "Comparing Corpora Using Frequency Profiling". In *Proceedings of the Workshop on Comparing Corpora, ACL 2000*. Hong Kong. 1-6.

Schmid, H. (1994). "Probabilistic Part--of--Speech Tagging Using Decision Trees". In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK.