A lexicographic appraisal of an automatic approach for detecting new word-senses

Paul Cook,[♣] Jey Han Lau,[♣] Michael Rundell,[♦] Diana McCarthy,[♠] and Timothy Baldwin[♣]

♣ Department of Computing and Information Systems, The University of Melbourne
♦ Lexicography MasterClass and Macmillan Dictionaries
♠ Department of Theoretical and Applied Linguistics, University of Cambridge
Email: paulcook@unimelb.edu.au, jeyhan.lau@gmail.com,
michael.rundell@lexmasterclass.com, diana@dianamccarthy.co.uk, tb@ldwin.net

Abstract

Over the last 20 or so years, lexicographical tasks, such as finding collocations and selecting examples, have been automated to some degree, both supplementing lexicographers' intuitions with empirical data, and reducing the "drudgery" of lexicography to allow lexicographers to focus on tasks which cannot easily be automated. Automated determination of word senses and identification of usages of a given sense, however, have proven difficult due to their covert nature. In this paper, we present a method, based on an automatic word sense induction system, for identifying novel word senses in a more recent Focus Corpus with respect to an older Reference Corpus. We evaluate this method in the context of updating a dictionary, and find that it could be a useful lexicographical tool for identifying new senses, and also dictionary entries whose definitions or examples should be updated.

Keywords: computational lexicography, neologisms, word senses, word sense induction

1. Updating dictionaries

Lexicography is expensive. Despite the falling cost of corpus resources, the process of compiling and editing dictionary text remains labour-intensive. This applies not only to developing new resources from scratch, but also to the (more usual) job of updating existing dictionaries. One promising strategy for publishers is to automate some of the editorial tasks, and significant progress has been made in this area over the last ten years (Kilgarriff and Rychlý, 2010; Rundell and Kilgarriff, 2011; Rundell, 2012). In brief, corpus-analysis software can aid in: (1) determination of the syntactic, collocational, and text-type preferences of a given word or meaning; (2) selection of a shortlist of suitable example sentences; and (3) (at a later stage) streamlining of the process of editing and finalising dictionary text. The current approach to dictionary development has the software presenting data to the lexicographer in a useful predigested form. But recent advances offer the prospect of a model where "the software selects what it believes to be relevant data and actually populates the appropriate fields in the dictionary database" (Rundell and Kilgarriff, 2011, page 278), leaving the human expert to validate (or refine, or reject) decisions made by the computer.

The various components of this model have all been trialled on real dictionary projects, providing the conditions for incremental improvements in performance. The GDEX software, for instance, which automatically finds appropriate dictionary examples in a corpus, was used initially on a project at Macmillan, when there was a requirement for a large number of new example sentences for specific collocational pairs (Kilgarriff et al., 2008). The results were uneven but broadly positive, with the editorial team completing the task more quickly than if they had taken a purely "manual" route. Versions of GDEX have since been used in other ventures. The heuristics and weightings have been optimised for a number of languages (e.g., Kosem et al., 2011), and the software is now a standard feature in the editorial toolkit of a number of dictionary developers.

In other areas, progress towards automation has been slower. But the direction of travel is clear: we are gradually putting together a suite of robust applications which collectively streamline the job of compiling and editing dictionary text. If the effect of all this is to transfer some lexicographic tasks from humans to machines, the goal is to produce better dictionaries at a lower cost. A striking outcome of the work done so far in this area is that automation not only delivers efficiency savings but also leads to improvements in quality. Automating a process forces us to go back to first principles and be explicit about what the task involves. What, for instance, are the features of a "good" dictionary example, or at what point can we say with confidence that a particular syntactic pattern is "typical" of a word? All of which is contributing to the goal of producing dictionaries that are more systematic, more internally-consistent, and less reliant on the subjective judgment of individual lexicographers.

Improving the language-description process presupposes having some language that needs describing. Methodologies for extracting candidate headword lists from corpora are already well-established. Meanwhile, the requirement for tracking language change (more pressing than ever now that most dictionaries are online and their users expect them to be up-to-date) is also being addressed, and the task of identifying emerging new words is benefitting from computational approaches (Rundell and Kilgarriff, 2011, pages 263–267). But (notwithstanding the media's obsession with shiny new headwords), there is more to updating a dictionary than adding neologisms. Two other salient aspects of keeping a dictionary up-to-date are finding novel senses of existing words, and ensuring that dictionary entries reflect contemporary conditions and technologies.

From the 1980s, as computer technology moved out of its specialist ghetto to become part of most people's everyday experience, words like *mouse*, *icon*, *virus* and *window* acquired new senses. (The word *computer* itself, for that matter, began life in the 17th century as a job title for someone whose work involved calculation.) Earlier dictionaries do not include these meanings, so they had to be added. More recent examples include words like *cloud* and *tablet*, *hybrid* (a type of car), *sick* (used in contemporary slang as a term of approval), and *toxic* (when referring to financial

assets or debts). None of these meanings existed when the Macmillan Dictionary was first published (in print form) in 2002, and all have been added to the online edition (Macmillan English Dictionary Online, hereafter MEDO).[1] An equally important, but more elusive, goal is to ensure that definitions and examples reflect contemporary realities. In recent updates to MEDO, for example, changes have been made to the definitions of *meeting* (participants do not have to be in the same location), *marriage* (not just between a man and woman), and indeed *dictionary* (no longer simply "a book which ..."). MEDO has also targeted example sentences with dated contexts, like this one exemplifying one of the meanings of the verb *to slot*:

(1) *She slotted another tape into the cassette player.*

Traditionally, these are labour-intensive operations. In an ideal world, a well-funded editorial team would carefully review every entry, consulting contemporary corpus data, and identify anything that needed changing or updating. This is increasingly impracticable. Budget constraints weigh heavily on most non-commercial institutions, while commercial lexicography is in the process of replacing a simple and reliable business model (selling books) with something more complex and (for the time being) less profitable.

So, for the sake of both systematicity and feasibility within limited budgets, it makes sense to see how far we can automate the tasks of finding novel senses and identifying other areas of the text that might need updating.

In this paper, we examine a previously-proposed technique for automatically identifying word senses that are new to one corpus with respect to another (Lau et al., 2012), based on an automatic word sense induction system. We propose a further extension to that system which can incorporate human intuitions about topics for which we expect to see many new word-senses. We describe our previous evaluations of the core system, and its ability to identify new word-senses. We then present a new evaluation of our proposed method in the context of updating a dictionary, in collaboration with a professional lexicographer (the third-named author of this paper). Our findings suggest that this method could indeed be a useful new addition to the lexicographer's toolkit.

2. Automatic novel sense detection

Word sense induction (WSI) is the task of automatically grouping the usages of a given word in a corpus according to sense, such that all usages exhibiting a particular sense are in the same group, and each group includes usages corresponding to only one sense (Navigli, 2009). The category "word sense" is not of course uncontroversial. There is no general agreement about what constitutes a discrete

[1] http://www.macmillandictionary.com/

meaning of a word, and dictionaries often exhibit considerable variation in their treatment of the same polysemous word. But although word meanings are unstable entities, often with shifting boundaries, dictionary conventions traditionally require that lemmas are divided up into numbered senses, and a good lexicographers' style guide will provide criteria for doing this.[2] Here, we describe a WSI technique we developed and its application to the task of identifying novel word senses.

The WSI methodology we use is based on a model we previously proposed (Lau et al., 2012). The core machinery of this method is driven by probabilistic topic models (Latent Dirichlet Allocation, LDA: Blei et al.,2003), where latent or unseen topics are viewed as the driving force for generating the words in text documents. In this model a document is viewed as a probability distribution over topics, and each topic is represented as a probability distribution over words. The probability distributions for documents and topics are automatically "learned" from the corpus. Crucially, the "topics" in a topic model do not necessarily correspond to topics in the sense of the subject of a text. Applying topic models to induce the word senses of a lemma of interest, these "topics" are interpreted as the induced senses.

In traditional topic models, the number of topics to be learnt is a parameter that must be set manually in advance. In WSI, this parameter translates to the number of senses to be induced for a lemma. To develop a model without this requirement, and which can learn varying numbers of senses for different lemmas as appropriate, we used a Hierarchical Dirichlet Process (HDP, Teh et al., 2006), a variant of LDA that also learns an appropriate number of topics/senses.

Following our previous work, for each usage of a target lemma we extract a three-sentence context, where the second sentence contains the usage of the lemma, and the first and third sentences are the preceding and succeeding sentences, respectively. These three-sentence snippets are viewed as the "documents" in the topic model. We represent each document as the bag-of-words it contains, as is common for topic models.[3] We also include additional positional word information to represent the local context of the target lemma. Specifically, we introduce an additional word feature for each of the three words to the left and right of the target lemma. An example of the features is given in Table 1. To illustrate the senses induced by our model and the usages that correspond to the senses, we present Tables 2 and 3 respectively, for the example lemma *cheat*.

[2] For a full discussion of word senses, see Hanks (2013, pages 65–83).

[3] We use the term bag-of-words to refer to the multiset of items occurring in some context, as it is commonly used in natural language processing. As described in Sections 3 and 4.1, we lemmatise our corpora. Our "bag-of-words" representation is therefore in fact a bag-of-lemmas.

Target lemma	dog
Context sentence	Most breeds of dogs are at most a few hundred years old
Bag-of-word features	most, breed, of, be, at, most, a, few, hundred, year, old
Positional word features	most_#−3, breed_#−2, of_#−1, be_#+1, at_#+2, most_#+3

Table 1: An example of the topic model features.

Sense Number	**Top-10 Terms**
1	heat think want ... love feel tell guy include find
2	cheat student cheating test game school to teacher exam study
3	husband wife cheat wife_#1 tiger husband_#-1 on ... woman marriage
4	cheat woman relationship cheating partner reason man woman_#-1 to spouse
5	cheat game play player cheating poker to card cheated money
6	cheat exchange china chinese foreign cheat_#-2 cheat_#2 china_#-1 to team
7	tina bette kirk walk accuse mon pok symkyn nick star
8	fat jones ashley pen body taste weight expectation parent able
9	euro goal luck fair france irish single 2000 point complain

Table 2: The top 10 terms for each of the senses induced for the lemma cheat.

Sense number	**Usage**
4	While I was single I slept with several married men. I had relationship with them. Now that I am married I feel horrible for having done so. I am always afraid my husband is going to cheat on me.
	It appears to me that there are people who are just disloyal. A man who cheats on his wife will cheat other partners whether that partner is a business partner or a wife – disloyalty transfer.
	I find it ignorant when men cheat on their wife, and when they found out the wife was sleeping around, they get mad. That makes no sense.
5	Lastly, the foremost argument in my personal opinion is that the profit margin of the online poker room is so large, that they simply would not need to cheat their own players. They are practically doing it already. Fairly.
	Do you feel you have been cheated when playing online poker? Well, guess what. You have been! The question is: do you want to continue being cheated?
	“There is not a card player who would not cheat, if he knows how.” - Walter Irving Scott, the phantom of the card table.

Table 3: Corresponding usages for induced senses 4 and 5 of the lemma cheat.

To identify novel senses, we compare a Focus Corpus with a Reference Corpus. In the application we consider here (updating a dictionary), the Focus Corpus would consist of newer texts; the Reference Corpus, on the other hand, would be older material, and common usages in this corpus would be expected to be reflected in the dictionary. (Details of the Reference and Focus Corpora used in this study are given in Section 4.1.) We combine the Focus and Reference Corpora to produce a supercorpus. For a given lemma of interest we then apply our WSI methodology to all of its usages in this supercorpus. (In this study we consider all lemmas meeting some frequency and keywordness cutoffs, also described in Section 4.1.) The WSI step automatically labels each usage of the lemma with its induced sense. We then calculate the "novelty" of an induced sense in the Focus Corpus as the ratio of its relative frequency in the Focus and Reference Corpora, akin to a simple approach to keywords (Kilgarriff, 2009), but applied to induced senses. We rank the lemmas according to the novelty of their highest-scoring induced sense. The highest-scoring induced sense for a given lemma is referred to as its novel sense.

New senses often arise for prominent cultural concepts (Ayto, 2006). In this paper, we introduce a new variant to our method for identifying novel senses that incorporates this observation. We first manually form a list of terms related to a particular topic (computing and the Internet for the analysis presented in Sections 4 and 5). For each induced sense we then determine its relevance to this topic based on its probability distribution over words from the topic modeller. We independently rank each induced sense by its relevance and its novelty score, and then rank each induced sense by the sum of its rank under each of these two rankings. This approach identifies induced senses which are both novel and related to a particular topic, and is referred to as "rank sum".

3. Previous evaluation

In this section we describe previously-presented evaluations of the WSI component of our method on several benchmarked WSI tasks, and an evaluation of the accuracy of our method for detecting whether a given word exhibits a novel sense in a more recent Focus Corpus compared to an older Reference Corpus and, furthermore, whether it can detect specific instances of a novel sense within the Focus Corpus. In Sections 4 and 5 we present a new evaluation of our method for identifying novel senses in the context of updating a dictionary.

Our WSI technique was first presented in Lau et al. (2012), and was initially evaluated using two datasets (Agirre and Soroa, 2007; Manandhar et al., 2010) to compare the system to the state-of-the-art in WSI. These datasets were produced within the auspices of a series of international events (SemEval, formerly SENSEVAL) for the objective comparison of computational systems that provide semantic analysis. Both datasets require the systems to induce senses for a sample of lemmas from some

training data and then label some unseen data with these senses. From the evaluation, our system outperformed the state-of-the-art systems, given the same conditions for tuning parameters. Moreover, on the more recent 2010 dataset our model, which uses HDP to automatically learn the optimal number of topics (senses), outperformed a more basic LDA model even when the latter was manually told how many topics to learn.

More recently we evaluated our WSI technique by participating in two SemEval 2013 WSI tasks. "Word Sense Induction for Graded and Non-Graded Senses" (Jurgens and Klapaftis, 2013) was similar to the previous WSI evaluations considered, but additionally required systems to identify not just the single most appropriate induced sense for a given test usage, but rather all applicable senses, and the extent to which they apply. In this evaluation a number of different metrics were considered, with our method outperforming all other participating systems in terms of one metric, and achieving strong results overall (Lau et al., 2013a). "Evaluating Word Sense Induction & Disambiguation within an End-User Application" (Navigli and Vannella, 2013) considered whether WSI can be applied to diversify search engine results. In this task our system performed best out of all participating systems, further demonstrating the effectiveness of our WSI approach (Lau et al., 2013b).

To evaluate the application of our WSI method for novel sense detection, our earlier work (Lau et al., 2012) provided the first, and to date only, available dataset, albeit a relatively small one. The production of such a dataset is difficult because word senses are covert and manually labelling occurrences in a corpus is a very time-consuming and laborious process. We focused on a small sample of lemmas which were identified as having senses arising in the period between the early nineties and 2007. This period was selected simply because of the availability of a Reference Corpus, the British National Corpus (BNC, Burnard, 1995), and a more recent Focus Corpus, the ukWaC (Ferraresi et al., 2008), produced automatically from data from the Web in 2007.[4] Since these corpora are of different sizes, they were made more comparable by using only the written portion of the BNC and extracting a similar-sized random sample of documents from the ukWaC and using TreeTagger (Schmid, 1994) to tokenise and lemmatise both corpora.

We used the Concise Oxford English Dictionary editions which best reflected contemporary usage for the two respective time periods: Thompson (1995, COD95) and Soanes and Stevenson (2008, COD08). Working on the assumption that new senses often arise for culturally salient concepts (Ayto, 2006), we directed our search towards entries relevant to computing and with sufficient frequency (more than 1000) in the BNC. The lexical selection was supported with a manual inspection of 100 random occurrences from the respective corpora and also a manual inspection of the

[4] Note that the new evaluation presented in this paper uses different Reference and Focus Corpora than our earlier work.

collocates of the candidate lexemes using word sketches (Kilgarriff and Tugwell, 2002).[5]

The above procedure yielded five genuine lemmas with a novel sense arising in the respective period.[6] We then selected five distractor lemmas with the same part of speech as a target and of similar frequency within the BNC, but where there was no evidence of a new sense given the respective entries in COD95 and COD08. The automatic WSI method was applied to the similarly-sized set of the documents from the BNC and the ukWaC and the output used for ranking the lexical items by their novelty score. The lemmas with a high novelty score had significantly higher ranks compared to the distractors; meanwhile, a baseline which only considered the frequency difference across the two corpora did not produce a significant difference in ranking. We additionally used the manually tagged samples to demonstrate that not only could the approach successfully rank lemmas on the basis of novelty, but also it could be used to identify the novel occurrences in the Focus Corpus. Promising results were obtained overall simply by identifying the specific novel sense with the topic that was automatically ranked highest for novelty and using that to identify occurrences. Furthermore, because the induced senses are modelled as lists of salient words, topic models afford a readily interpretable representation for word sense, highlighting the potential for such automatic methods to produce output that can inform the lexicographic process.

4. Lexicographical evaluation

In this section we describe an evaluation of our proposed method for identifying novel word senses in the context of updating a dictionary, based on manual analysis by a lexicographer.

4.1 Corpora and pre-processing

Our previous evaluation of the ability of our WSI method to identify novel senses (presented in Section 3) used the BNC and ukWaC, corpora which consist of very different genres. For this analysis we consider more-comparable corpora. We use the English Gigaword Fourth Edition (Parker et al., 2009), henceforth referred to as GIGAWORD, which consists of newswire articles from six services including the New York Times Newswire Service; the Los Angeles Times/Washington Post Newswire Service; and the Agence France-Press, English Service for the years 1994–2008.[7] For our Reference and Focus Corpora we use the sub-corpora of Gigaword for the years 1995 and 2008, respectively, the earliest and latest years in the corpus for

[5] http://www.sketchengine.co.uk/

[6] The five lemmas were *domain* (n), *export* (v), *mirror* (n), *worm* (n), and *poster* (n).

[7] There is a fifth edition of this corpus which additionally includes data for 2009 and 2010, but we unfortunately do not have a license for this edition of the corpus.

which data from all services are available. This provides Reference and Focus Corpora which are comparable, in that they both consist of newswire data from the same sources for a given year, although there are of course topical differences between the corpora for the two years. Moreover, these corpora are diverse, consisting of data from six sources, although all data are from newswires.

Gigaword consists of several document types with the by far most frequent being "story", which corresponds to a typical newswire story. We only consider these documents. Gigaword is known to contain a substantial number of Spanish documents. To reduce the amount of non-English content in our corpora, we filter all documents not identified as English using langid.py (Lui and Baldwin, 2012), a statistical language identification tool. Newswire text contains duplicate and near-duplicate documents, corresponding to, for example, an update to a previous story. We apply exact deduplication, and near-deduplication using Onion (Pomikalek, 2011), to remove such documents. Finally, we part-of-speech tag and lemmatise the resulting corpora with TreeTagger (Schmid, 1994), in line with our earlier experiments over the BNC and ukWaC.

The Reference (1995) and Focus (2008) Corpora consist of 193M and 202M words, and 471k and 536k documents, respectively. We count the words in each corpus, and compute keywords using the method recommended by Kilgarriff (2009). We identify all nouns with frequency greater than 1000 in each corpus, frequency less than that of the 100th most-frequent noun in each corpus, and keywordness between 0.5 and 2. This gives 3185 nouns over which we run our proposed method for identifying novel word senses.

For the "relevance" component of the rank sum method for identifying novel word-senses we manually identify words related to computing and the Internet, topics that increased in prominence between the time periods of our Reference and Focus Corpora. We compute keywords for our Focus Corpus relative to our Reference Corpus, again using the method of Kilgarriff (2009). This method includes a parameter, α, which roughly controls the frequency range of the resulting keywords. We identify the top-1000 lower-case keywords with length at least three for α set to 1, 10, and 100 to consider keywords with a range of frequencies. The first and second authors of this paper independently annotated the keyword list to identify those that they judged to be primarily related to computing and the Internet in the newswire domain. Thirty-three keywords were selected by both annotators, and these words were used as the domain-specific words in computing relevance.

4.2 Lemma selection

We ran our method for identifying novel word senses on all 3185 nouns matching our frequency and keywordness criteria from the previous subsection. We considered both the novelty and rank sum approaches. The top-10 items for each method were selected for further analysis.

It is possible that our proposed method fails to identify many new word-senses, i.e., that amongst the lemmas not identified by our system there are many new senses. In an effort to evaluate this we also analysed ten randomly selected lemmas. The thirty lemmas analysed are shown in Table 4.

Novelty	Rank Sum	Random
airstrikes	**advertiser**	arena
candy	**cell**	**audit**
cleric	**click**	beauty
junta	**copyright**	**follow-up**
militiaman	**fingerprint**	**fraction**
nutrition	**instinct**	likelihood
plastic	**search**	lyric
prostitution	**text**	stockpile
truce	**video**	taxis
vest	**web**	tension

Table 4: The 30 lemmas selected for analysis and the method through which they were selected (presented in alphabetical order in each column). For items shown in **bold-face** the analysis revealed a noteworthy change in usage.

4.3 Analysis process

For each lemma we produced a summary consisting of the following information:

- The words associated with the topic corresponding to the candidate novel sense (provided by the topic modeller);
- The ten highest confidence novel sense usages from each corpus;
- The number and proportion of usages corresponding to the novel sense in each corpus;
- A random sample of ten usages from each corpus.

These summaries were then given to a professional lexicographer to analyse. Crucially, the lexicographer (the third-named author of this paper) did not know whether a given lemma was included because it scored highly for the novelty or rank sum method, or because it was one of the randomly selected items. The analysis was carried out with respect to the following questions.

- Would the candidate novel sense be included in various types of dictionaries (e.g. a general pedagogical dictionary, a large “native-speaker” dictionary, an online dictionary)?
- Has the candidate novel sense already been included in dictionaries, but only in those for specialised domains?
- Is the candidate novel sense interesting for some other reason?

Throughout the analysis two "reference" dictionaries were consulted:

MEDO Macmillan English Dictionary Online: a medium-sized, monolingual, mainly pedagogical dictionary with approximately 50,000 headwords;[8]

ODE Oxford Dictionary of English: a standard monolingual "desktop" dictionary aimed at native speakers with about 80,000–90,000 headwords.[9]

Two other dictionaries aimed at a similar market to MEDO were also sometimes referred to: the Cambridge Advanced Learners Dictionary (CALD),[10] and the Longman Dictionary of Contemporary English (LDOCE).[11]

5. Analysis

Table 4 shows the lemmas analysed, displaying which were found to have a notable difference in usage between the Reference and Focus Corpora. Overall there are more "interesting" findings for the lemmas obtained through novelty than the randomly selected lemmas. Moreover, for the rank sum method, all lemmas correspond to an interesting difference in usage in the Focus Corpus. This suggests that our proposed method could be a useful tool for identifying changes in usage.

5.1 Uninteresting findings

For the lemmas not shown in boldface in Table 4, a notable difference in usage was not observed between the Reference and Focus Corpora. In all of these cases the data provide no evidence of a novel sense in the Focus Corpus, and the sense instantiated in the data is adequately covered in the two "reference" dictionaries considered, and in other general dictionaries. For the "random" lemmas this is not surprising, and we will not discuss them further here.

In the case of each item in this category identified by novelty (i.e., *airstrikes*, *candy*, *junta*, *plastic*, *prostitution*) there are marked contextual differences between the new and old corpora, and random and selected sets of usages. Here the proposed method has identified a novel configuration of frequent collocates — a sudden spike which typically reflects a (briefly) salient news story. Thus at *junta*, the collocates list (including *myanmar*, *aid*, *cyclone*, *relief*) relate to a cyclone which hit Myanmar/Burma in 2008,[12] causing huge loss of life. Similarly, the data for *candy* in the Focus Corpus are skewed by a news story about Chinese candy being contaminated by melamine. What tends to happen in these cases is that the other data

[8] http://www.macmillandictionary.com/

[9] http://oxforddictionaries.com/

[10] http://dictionary.cambridge.org/

[11] http://www.ldoceonline.com/

[12] http://en.wikipedia.org/wiki/Cyclone_Nargis

(selected data from the Reference Corpus and all random data) exhibit the same sense but a wider range of contexts. Topical differences are known to be challenging for methods for identifying differences in word sense between corpora (Peirsman et al., 2010), and indeed similar observations in our earlier work led to the development of the rank sum method to address this. That none of the top-10 lemmas for the rank sum approach are in this category suggests that it has been successful in this regard.

5.2 Dictionary account needs tweaking

In the following cases, the data provide evidence which suggests that some existing dictionary accounts (sometimes in MEDO, sometimes in the other dictionaries referred to in Section 4.3) may need to be tweaked or broadened. Most of these cases, however, do not indicate the emergence of a genuine new word-sense.

advertiser Examples from the Focus Corpus refer overwhelmingly to web advertising—but many of those from the Reference Corpus do too. Web advertising was already established in 1995, and MEDO's entry reflects this (though that is not the case in some other dictionaries). Several of the corpus examples for *advertiser* include references to *publishers*, and many dictionaries are still lagging in their definitions of what "publishing" entails (typically focussing on the traditional media of books, music, journals, and the like). So the co-occurrence of *advertiser* and *publisher* in the data serves as a useful reminder that one or both of these entries may need updating to reflect the words' contemporary use.

cell In both the Focus and Reference Corpora, the examples refer to *cell phones* (the usual term in American English, though not in British English, where *mobile (phone)* is preferred). All the dictionaries examined record this use of *cell*. However, the Focus Corpus includes at least two references to *cell sites*, and this appears to be a valid term, defined in Wikipedia as: "a cellular telephone site where antennas and electronic communications equipment are placed". *Cell site* does not appear in any general English dictionary, but it is at least worth considering whether it should.

cleric The data from the Focus Corpus overwhelmingly refer to *Muslim clerics* (who are typically characterized as *radical* and/or *fundamentalist*), and this marks a clear shift from what we find in the Reference Corpus, where *cleric* tends to suggest an innocuous Church of England figure of the type found in a Trollope novel. Although the entries in the two "reference dictionaries" both take account of this change, the definitions and/or example sentences in some dictionaries do not: LDOCE, for example, defines *cleric* simply as "a member of the clergy".

copyright The Focus Corpus data often mention *copyright* in the context of new media (games or software, for example), whereas older data refer to more traditional contexts (songs, books etc.). There is no change in the essential meaning ("protection of ones' intellectual property"), but some dictionaries may need to update definitions and/or example sentences in order to account for the broader scope of this term.

militiaman The data suggest that some updating is required at dictionary entries for *militia*. (*Militiaman* itself is adequately defined as "a member of a militia".) The Focus Corpus contexts point to the now dominant use of *militia* to refer to an unofficial

armed group, typically with links to terrorism or insurgency (*Shiite militias*, etc.). The current definition in MEDO ("a group of ordinary people who are trained as soldiers to fight in an emergency") invokes an older, more neutral use, referring to a citizen army, and most other dictionaries have the same emphasis.

truce The selected examples (Focus and Reference) reflect the standard use of *truce* (and the contexts – mostly to do with Palestine and Israel – show depressingly little change over the period). But the randomly selected usages include at least two cases where the context is not war, but business or politics. This may indicate a separate sense: more a cessation of argument or opposition than of fighting and hostilities. The current definition in MEDO could be said to cover all these scenarios: "an agreement between two people or groups involved in a war, fight, or disagreement to stop it for a period of time". But the example sentences all refer to war-type contexts and ODE's entry has a similar focus.

vest All the Focus Corpus examples (but only one or two from the Reference Corpus) refer to "suicide vests" or "explosive vests" – evidently a salient context in contemporary texts. The closest sense in MEDO defines a vest as "a piece of clothing with no sleeves or collar worn over other clothes, for example for protection", and follows with an example: *a bulletproof vest*. This does not fully reflect current usage, so the entry may need tweaking.

video The Reference and Focus Corpora have very different emphases, with the newer data referring exclusively to online videos (with collocates such as *circulate*), whereas the older data refer to movies or TV programmes stored on VHS devices (the prevailing technology in the 1990s). Here, the entire MEDO entry is out of date (it refers to material "recorded on videotape") and had in fact already been flagged for attention in the next update. ODE has already updated its entry to take account of changing technologies, and its definition reads: "a recording of moving visual images made digitally or on videotape". This is not a novel sense as such, but the dictionary record definitely needs updating.

web In the sense of "the Web", this is a fairly recent but by no means novel meaning. One interesting point is that the Reference Corpus data include several citations for the expression *world wide web*, which is now very dated. Most dictionaries have a neutral entry for this term, and in many (including ODE and MEDO), definitions of *web* or *the Web* simply say "the World Wide Web", cross-referring to another entry. In 2013, this is the wrong way around – rather like defining *bus* as "an omnibus" (as would have happened in dictionaries 100 years ago). So here again the data serve as a useful reminder to make adjustments to an entry which could easily have been ignored.

Two of the "random" lemmas – *follow-up* and *fraction* – were also assigned to this category when the data were analysed from a lexicographic viewpoint. It is not so surprising that a randomly chosen lemma would appear in different contexts, given the different dates of the two corpora. Since we have already established that the automated method tends to find more noteworthy cases than random ones, we do not discuss the "random" lemmas further here.

5.3 Novel senses

For these lemmas, the data indicate a genuine novel sense.

click The use of *click* meaning "an instance of a user clicking on something" was already established in 1995. MEDO includes this meaning, with the example: *You can order anything with a single click*. However, examples like the following suggest a newer use:

(2) *Total paid clicks in the fourth quarter rose 30 percent from the same 2006 period.*

(3) *For instance, comScore estimated Google's fourth-quarter clicks increased 25 percent.*

This reflects the Web business model, where each click on an advertising link represents a specific value for the publisher. The current MEDO entry does not adequately cover this newer use, which (though more specialised) is nevertheless valid.

fingerprint Several examples from the Focus Corpus data refer to *digital fingerprint* (which is not found in the Reference Corpus data). Most likely this simply refers to a digital record of a fingerprint. But the term *digital fingerprint* is also used in data security contexts with a different meaning. This second meaning does not appear in any of the four general dictionaries we consulted (see Section 4.3). But it is recorded in the more specialised businessdictionary.com, where it is defined as: "Coded string of binary digits (generated by a mathematical algorithm) that uniquely identifies a data file". This is followed by the more familiar second sense: "Analog fingerprint of a person converted (digitised) into a binary file". There may be a case for a similar two-sense entry in general dictionaries.

search The Focus Corpus provides evidence for a novel sense of *search*, and this is absent from the Reference Corpus. The novel sense refers to the business of search (on the Web), and is an uncountable noun (distinct from "doing a Google search for something"). This use was added to MEDO in an update carried out in early 2013, as follows: "3 [uncountable] the process of searching for information on the Internet, or the business and technology that supports this": *Founded in 1995, Yahoo was quick to get into search*. This use is not currently accounted for in most dictionaries.

text All the data from the Focus Corpus relate to *text messaging*, which was still rare in 1995 and does not appear in the data from the Reference Corpus. (The BNC has no examples of *text messaging* either.) There has clearly been a huge shift in the frequency profile of the word *text* over this period. The proposed automated method has successfully identified this newer usage, though in this case it is something that all the checked dictionaries take account of.

One of the "random" lemmas, *audit*, was also found to exhibit a genuine novel sense in the data considered here. The Focus Corpus usages refer mainly to the contexts of aviation and slaughterhouses, and indicate an inspection aimed at ensuring safety and compliance with regulations. This appears to be fairly recent (the Reference Corpus data – both random and selected – focus on the older "financial audit" sense, the work done by *auditors*). What we see here is probably a fairly recent sense, though there is some evidence for it in the (1992) BNC, e.g. for *environmental audit* (43 hits), and most dictionaries already cover it.

5.4 Other cases

instinct The data from the Focus Corpus relate to a smartphone with this proprietary name released in 2008 (hence collocates like *iphone* and *samsung*), and the word always appears with initial uppercase (*Instinct*). This usage would not typically be recorded in the dictionaries consulted in this analysis, and it could potentially be identified by more simple means (such as a keyword analysis in which case is preserved). However, information about case is not available to the automated method, and so from this limited perspective, the system has successfully identified that *instinct* has a new usage in the Focus Corpus.

nutrition All selected examples of *nutrition* in the Focus Corpus are of the following type:

(4) *NUTRITION Per serving (based on 8): 179 calories, 2 g protein, 42 g carbohydrates, 1 g fat, 0 g saturated fat, 0 mg cholesterol, 67 mg sodium, 5 g dietary fiber*

This relates to a standard format for nutritional information on food labelling. Although this cannot be considered a novel sense, it is a usage which is far more common in the Focus Corpus than in the Reference Corpus, and the proposed method has identified it.

6. Conclusions

We presented an automatic method for identifying new word-senses in a Focus Corpus of more recent texts with respect to an older Reference Corpus. An evaluation of our method in the context of updating a dictionary suggests that this method has promise as a tool for helping lexicographers to identify new word-senses. Moreover, this method was shown to have the potential to aid in identifying dictionary entries that require updating, for example, because definitions or example sentences are out of date. Crucially, although these tasks are important for keeping dictionaries current, they are also very expensive, and there have been few previous efforts to automate them.

At the heart of our proposed method is a word sense induction system, which groups together similar usages of a given word in a corpus. In future work we intend to consider whether this system can be applied to other dictionary writing tasks, for example, identifying good dictionary examples for a particular word sense, or semi-automatic dictionary drafting (Kilgarriff and Rychlý, 2010).

To encourage further research on topic modelling approaches (such as the one used by our system) in computational lexicography, and the use of our proposed method in lexicographical projects, we have made our word sense induction system publicly available under a license which permits its use for commercial purposes.[13]

[13] https://github.com/jhlau/hdp-wsi

7. References

Agirre, E. and Soroa, A. (2007). SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic.

Ayto, J. (2006). *Movers and Shakers: A Chronology of Words that Shaped our Age*. Oxford University Press, Oxford, UK.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Burnard, L. (1995). *User Guide for the British National Corpus*. Oxford University Computing Service, Oxford, UK.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, pages 47–54, Marrakech, Morocco.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, USA.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In Second Joint Conference on Lexical and *Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, USA.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress (EURALEX 2008)*, Barcelona, Spain.

Kilgarriff, A. and Rychlý, P. (2010). Semi-automatic dictionary drafting. In de Schryver, G.-M., editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*, pages 299–312. Menha Publishers, Kampala, Uganda.

Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In Corré ard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137. Euralex, Grenoble, France.

Kosem, I., Husak, M., and McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, pages 151–159, Bled, Slovenia.

Lau, J. H., Cook, P., and Baldwin, T. (2013a). unimelb: Topic modelling-based word sense induction. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311, Atlanta, USA.

Lau, J. H., Cook, P., and Baldwin, T. (2013b). unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the*

Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 217–221, Atlanta, USA.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

Manandhar, S., Klapaftis, I., Dligach, D., and Pradhan, S. (2010). SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). *English Gigaword Fourth Edition*. Linguistic Data Consortium, Philadelphia, USA.

Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masaryk University.

Rundell, M. (2012). The road to automated lexicography: an editor's viewpoint. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, pages 15–30. Oxford University Press, Oxford, UK.

Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In Meunier, F., Cock, S. D., Gilquin, G., and Paquot, M., editors, *A Taste for Corpora. In honour of Sylviane Granger*, pages 257–282. John Benjamins, Amsterdam, Netherlands.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Soanes, C. and Stevenson, A., editors (2008). *The Concise Oxford English Dictionary*. Oxford University Press, eleventh (revised) edition. Oxford Reference Online.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Thompson, D., editor (1995). *The Concise Oxford Dictionary of Current English*. Oxford University Press, Oxford, UK, ninth edition.