New Words Prioritization Engine: A System for Evaluating Multiple Data Inputs to Prioritize Neologisms for Inclusion in Dictionary Projects

Katherine Connor Martin

Oxford University Press

katherine.martin@oup.com

Abstract

With today's massive Web-based corpus resources, the key challenge facing lexicographers of new words in languages with a major digital presence is no longer *identification* of neologisms, but rather their *prioritization* for inclusion in the dictionary. There are many possible data points that can be leveraged to prioritize the most editorially significant from among tens of thousands of candidates, including frequency and diachronic trends in corpora, evidence of reader interest via Web searches, length of the word's documented existence, and salience of the item in particular regions, registers, or domains of editorial interest. The most effective way to use these data inputs is to take a holistic approach, considering multiple factors simultaneously. This paper discusses the use of a new system, Oxford's New Words Prioritization Engine, developed by Oxford Dictionaries to facilitate prioritization of large sets of candidate words by combining multiple sources of data in a single interface for analysis and by capturing human judgments about particular words so that they can be leveraged to improve future results.

Keywords

corpora, neologisms, prioritization, dictionary

Introduction

In the 21st century, as dictionaries have come to be primarily electronic resources, the print edition has been replaced by the online update as the vehicle by which new items are added. This

has accelerated both the rate at which new words are included, since major dictionaries may be updated multiple times per year, and the public profile of neologisms, since the new entries are often promoted in the media. Tracking and identification of neologisms, then, is a key task of ongoing dictionary projects, while limited editorial capacity has replaced limited page space as the primary constraint on adding new material.

Over the same period, increasing availability of digital resources for English and other well-resourced languages, including extremely large-scale Web-based corpora, has greatly facilitated neologism discovery. The dictionaries of the analogue age were built on a foundation of millions of hand-written index cards, any one of which might underpin an entire entry, but even the voluminous *Oxford English Dictionary* (OED) was not able to include all the words identified by the painstaking reading of texts (Burchfield 1973, 7). Now that lexicographers have access to multi-billion token corpora, the already daunting task of selecting from among candidates for full treatment in the dictionary has become even more difficult, and prioritizing from among the vast number of candidates for inclusion has become as important a consideration as identifying them in the first place.

Not all lexical items that would be new to a given dictionary will be new to the language as a whole; many new words represent ‘gaps’, in that they existed and could have been included in the dictionary previously but were not, either because they were overlooked or because they were deliberately excluded based on editorial criteria. For the purpose of the forthcoming discussion, *neologism* will be used loosely as synonym of *candidate word*, meaning any lexical item that would represent a new addition to the dictionary. For maintenance of a comprehensive lexical inventory, it is crucial that dictionary projects distribute their editorial resources to address gaps as well as genuinely new coinages. With infinite time and resources, one could

justify including every lexical item ever recorded, but with limited resources, we must distinguish those that are most useful or important. Prioritization in this context is the curation of editorial work to maximize value to dictionary users and to maintain balance among the various categories of candidates, such as general, specialist or technical, slang, and dialect terms.

This paper will first briefly discuss the opportunities and challenges of corpus-based neologism monitoring for dictionaries. Next it will introduce the problem of prioritizing among candidates with reference to the requirements of the OED and Oxford University Press's major current English monolingual dictionaries, the *Oxford Dictionary of English* (ODE) and the *New Oxford American Dictionary* (NOAD), which has motivated development of a new system at Oxford University Press (OUP), intended to facilitate prioritization of candidate words by aggregating data and capturing editorial judgments. It will then discuss the basic architecture of the system and its user interface, and present a case study in evaluating the efficacy and results of different candidate sources. Finally, it will discuss reception and feedback from users of the system, and plans for future development.

Practical Considerations of Corpus-based Neologism Tracking for Dictionaries

Large-scale monitor corpora and recent Web-based neologism-tracking systems like NeoCrawler (Kerremans et al. 2012) and Neoveille (Cartier 2017) have made it possible to identify emerging English neologisms at a very early stage, which facilitates broad study of neology as a phenomenon, but they are not optimized for the use cases of the practical lexicographer. Early identification of emerging neologisms is of obvious value to lexicographers, since the individual items identified, as well as broader trends, may inform further research for inclusion of new items in the dictionary. However, dictionaries tend not to add words until they have become reasonably well established in terms of both time and frequency, rendering the *hapax legomenon*

of lesser practical interest than the *myriakis legomenon* (*myriakis* = myriad, or literally, 10,000 times) that isn't yet covered in their particular dictionary.

Additionally, monitor corpora at the billion-plus token scale, such as BYU NOW corpus[1], the Josef Stefan Institute's timestamped corpora (JSI)[2], and OUP's monitor corpora, tend to be derived from a somewhat limited array of sources, primarily blogs and online newspapers. That may be an inherent limitation of a corpus aiming to capture reliably dated material from a genre-consistent set of sources over time at this scale, but it results in a skewed view of the lexicon, with some domains, such as business and sports, being relatively overrepresented by comparison with a more balanced corpus.

Table 1 compares the top ten nouns by frequency in COCA[3], a 520 million word genre-balanced monitor corpus, the OEC[4], a 2.5 billion token balanced corpus, and the JSI Timestamped Corpus 2014-2016 for English, a 21.3 billion token Web-based monitor corpus consisting mainly of news sources. The large size of the JSI corpus means that it is able to cast the largest net to identify and provide evidence of candidate words, but the high relative frequency of the nouns *game*, *team*, and *company* (by comparison with COCA and OEC) shows the impact of domain skewing. The same skewing by domain affects emerging neologisms, causing words in the overrepresented domains to appear to be far more common than they genuinely are relative to words outside those domains.

Table 1

COCA (520 million)	**OEC (2.5 billion)**	**JSI English 2014-16 (21.3 billion)**

[1] https://www.english-corpora.org/now/ (accessed April 17, 2019)
[2] https://www.sketchengine.eu/jsi-newsfeed-corpus/ (accessed April 17, 2019)
[3] https://www.english-corpora.org/coca/ (accessed April 17, 2019)
[4] https://en.oxforddictionaries.com/explore/oxford-english-corpus/ (accessed April 17, 2019)

lemma (noun)	freq.	lemma (noun)	freq.	lemma (noun)	freq.
year	769,254	*people*	3,842,914	*year*	50,548,353
time	764,657	*time*	3,648,568	*time*	33,350,185
people	691,468	*year*	3,459,144	*people*	28,892,785
way	470,401	*way*	2,233,873	*game*	19,760,539
day	432,773	*day*	1,929,139	*day*	19,334,818
man	409,760	*thing*	1,821,589	*company*	18,633,825
thing	400,724	*man*	1,627,994	*team*	17,529,452
woman	341,422	*life*	1,553,539	*way*	16,721,106
child	333,849	*part*	1,460,329	*week*	14,762,562
life	333,085	*work*	1,457,841	*state*	14,225,330

TOP TEN NOUNS BY RAW FREQUENCY IN THREE ENGLISH CORPORA

Furthermore, the lack of reliable computational solutions for identifying semantic neologisms (new senses) means that dictionaries must use other methods to track candidates belonging to this important category of lexical innovation.

Another peril of relying on monitor corpora for neologism identification is that the echo effect of the media, in which a single press release or announcement generates hundreds of news stories, amplifies the apparent frequency of certain topical items. As Brookes (2007) has noted, this phenomenon is particularly noticeable for scientific vocabulary.

For all of these reasons, while corpus-based neologism identification is an important tool for dictionaries, an update process relying on this type of analysis alone would be the equivalent of a one-legged stool; human identification of candidates still has a crucial role to play in a well-

rounded dictionary update program that reflects a broad range of registers, genres, dialects, and modes of expression. When an automated word-tracking system based on a monitor corpus was introduced to identify candidate words for the 11th edition of the *Chambers Dictionary*, the lexicography team also derived candidates from a manual reading program and a database of editorial suggestions (O'Donovan and O'Neill 2008). For today's online dictionaries, an additional source of neologism candidates must also be considered: Web or app searches for words that are not yet present, which can be seen as an indicator of the public's interest in particular words.

The Problem of Prioritization

For a mature dictionary that maintains a cycle of ongoing revision and expansion for digital publication, the availability of digital sources of neologism candidates results in an embarrassment of riches, as the burden of continuously reviewing multiple sources of candidate neologisms is immense, and can result in considerable duplication of effort if the same candidates are generated repeatedly over time or from multiple sources. At OUP, where the dozens of editors revising the historical OED work alongside the smaller team of lexicographers updating ODE and NOAD, a further complication is that candidates need to be evaluated for two different projects with distinct requirements.

As corpus-based dictionaries developed in the 1990s, ODE and NOAD already have robust coverage of contemporary English and their updates tend to focus on recent lexical innovations. However, these are not restricted to high-profile items of the sort that are promoted in media campaigns. In digital contexts, it is often advantageous to define fully items that might have been dismissed as deducible from their constituent elements in the print era, such as predictable derivatives, affixed forms, and compounds.

In contrast, the OED is undergoing a full revision for the first time in its century-long history and aims to fill gaps across 1000 years of English usage, so emerging 21st-century neologisms form a relatively small proportion of its identification and prioritization activities; "new words" added to the OED in the past year include not only 21st-century coinages like *stan* (2000) and *exomoon* (2008), but also items dating back to Middle English, such as *bedunged* (ante 1425).

Suggestions of new words submitted by in-house editors are approvvalidated for inclusion at a rate as high as 72% for the OED (Diamond 2016), but automatically identified candidates are less likely to be deemed draft-worthy. As editors experimented with new candidate sources to get a more comprehensive view on the lexicon, they often found that the chaff (words already covered, misspellings, non-words, trademarks and proper names) vastly overwhelmed the wheat.

To address this problem, a new tool, the New Words Prioritization Engine (NWPE), was developed at OUP to assist lexicographers in identifying and selecting new words. The system is designed to facilitate prioritization of candidates derived from any source by providing data from multiple types of evidence in a single interface for analysis. It seeks to combine human curation with automated scale, by leveraging lexicographers' instincts and experience to curate the collection of candidates and by capturing their judgments about particular words so that they can be leveraged to improve future results.

Description of the New Words Prioritization Engine

There are three major components of the NWPE system: ingestion of inputs, back-end processing (variant forms assessment and evidence gathering), and the Web-based user interface. In the initial phase of the system, the inputs and ranking criteria in the system are completely

variable and determined by users. To meet the needs of multiple editorial projects, the NWPE architecture doesn't exclude a candidate if it is already present in one dictionary (it might still be a valid candidate for another dictionary); instead, it assesses presence or absence in two separate dictionary databases (corresponding to the two lexicography teams that use the system) and retains that data in its output. These features are designed to make the system as flexible as possible for practical lexicography applications, including unforeseen use cases.

Inputs. NWPE inputs are lists of neologism candidates which are uploaded in the form of a spreadsheet. As of this writing, OUP lexicographers have uploaded 117 lists to NWPE, registering 91,551 distinct candidates. The input lists can be of any size and can be derived from any source the user wishes to analyze, but they tend to be large, computationally derived sets that are too time-consuming to assess manually because of high chaff:wheat ratios and frequent repetition. The following are examples of typical inputs:

- Corpus keyword analysis. The Word List function in Sketch Engine software is used to generate lists of keywords that are highly represented in a focus corpus or subcorpus relative to a reference corpus or subcorpus (Kilgariff 2009). This facility has been used by NWPE users to produce input lists of keywords that are particularly associated with documents from a specific geographic area (Nigeria, India, Ireland), specific subject domains (film), and specific time periods. Corpus keyword lists are used to ensure that the lexicon of a particular domain or region is treated consistently and are in this way an essential complement to targeted reading programs, which are more likely to generate a haphazard list of suggestions due to the smaller sample size.
- Trending words. The Trends function in Sketch Engine, which uses the Diacran framework (Kilgariff et al. 2015) to identify words undergoing significant change in

frequency over time, is used to generate lists of words that register a positive trend over monthly time-slices in OUP's Komodo monitor corpus.

- Morphological wordlists. A wordlist is extracted from a corpus based on a query to identify a particular type of formation (e.g., words with a particular suffix in a corpus derived from the Early English Books Online[5] corpus).
- Headword lists. Headword lists from subject reference dictionaries published by OUP have been ingested to identify missing words from a particular subject domain.
- Crowdsourced suggestions. Responses to appeals to the public for suggestions of particular types of vocabulary.
- Suggestions from reading programs. The OED operates a number of specialist reading programs by which experienced readers manually identify notable uses of new and existing words and record citations for them. The "Incomings" database in which the digital citations are stored contains more than 3 million quotations (Diamond 2016), each of which has one or more specific catchwords and a variety of document metadata. Lists can be extracted of catchwords corresponding to specific criteria, such as region, date, and subject domain.
- Failed searches. Lists of items searched unsuccessfully on OED.com or oxforddictionaries.com.

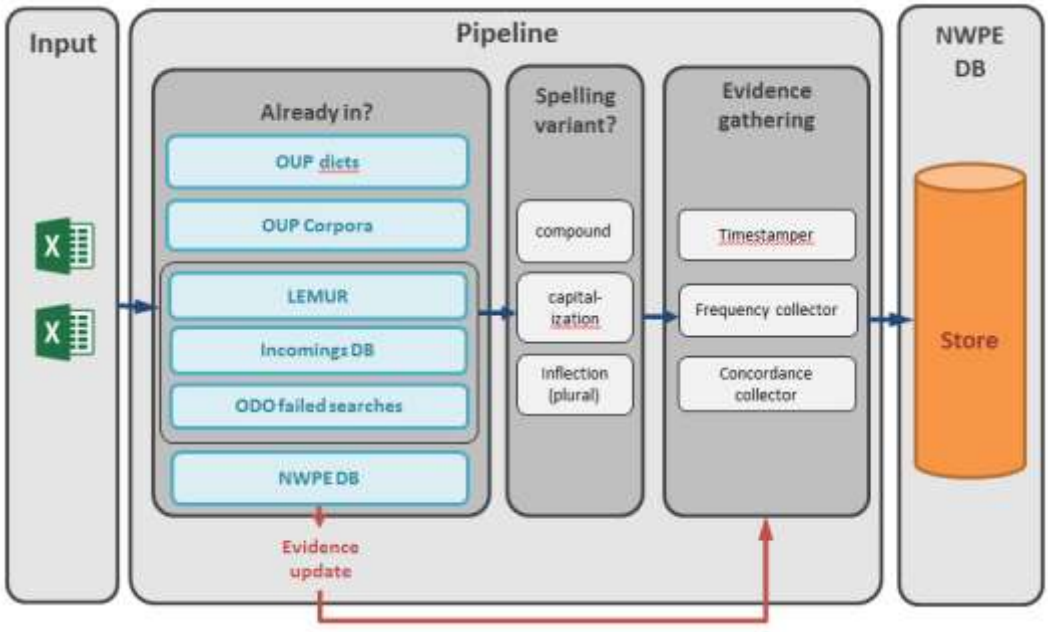
Sources like these had been used for candidate identification prior to the introduction of NWPE, typically in the form of discrete spreadsheets, but the new system makes it much more efficient to investigate them. Uploaded inputs are accessible and visible to all users, resulting in an

[5] https://www.textcreationpartnership.org/tcp-eebo/

accretion of information over time, and enabling items that are initially deprioritized to be revisited.

Back-end processing. In the first stage of processing, the wordform's presence or absence in a set of evidence sources is established. Next, a spelling variant assessment is performed on each wordform, which may be either a single word or a multi-word expression. The variant assessment extrapolates potential variant forms of the uploaded wordform based on inflection, hyphenation or spacing, and capitalization, in any combination. For instance, the ingested wordform "alternative facts" was found to have five different alternative forms registering at least one hit in NWPE's evidence corpora: "Alternative facts", "Alternative Facts", "alternative-facts", "alternativefacts", and "alternative fact" (singular noun).

Figure 1



COMPONENTS OF THE NWPE PIPELINE

Evidence is then gathered for the wordform and all of its predicted variants in the various evidence sources in which it is present. These evidence sources provide a number of different

types of information which can be used to assess the priority of the candidate words for drafting. The following data points are collected in the first iteration of NWPE:

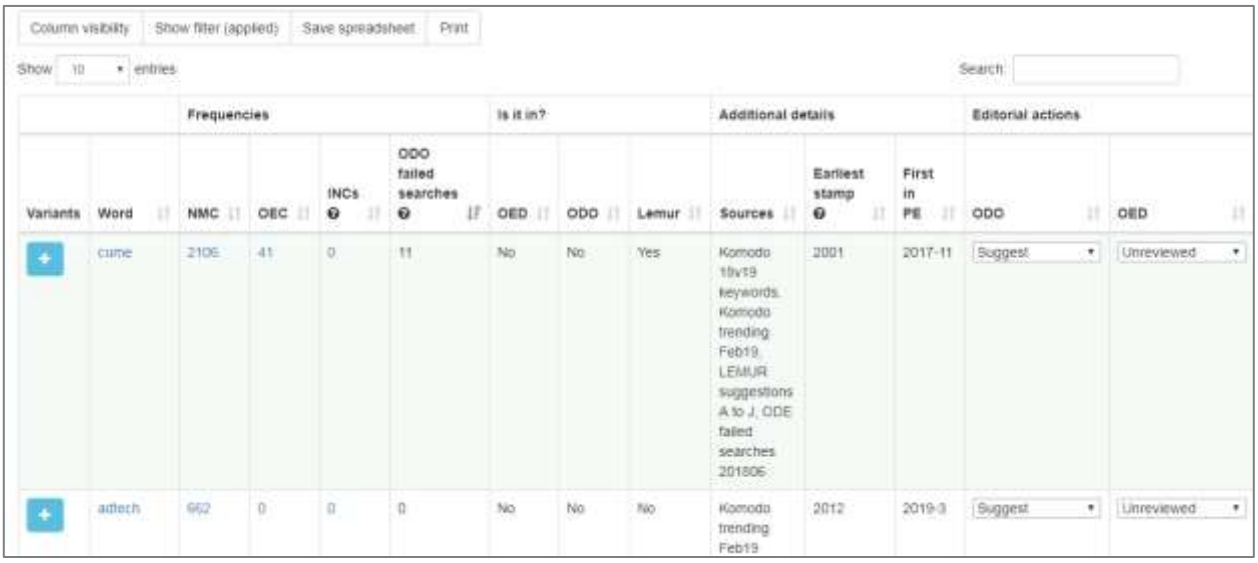
- Inclusion in reference dictionary databases. Each wordform is checked for exact matches against the lemma lists of the OED and the combined ODE & NOAD database.
- Inclusion in editorial suggestions database (LEMUR). Each wordform is checked for exact matches in LEMUR, the internal database of editor-generated suggestions identified either in the course of editorial work or via daily life. The LEMUR database currently holds over 40,000 suggestions; cross-referencing these with the data in NWPE assists in prioritizing from among these many items, and is particularly helpful in identifying items that have become significantly more common since they were originally suggested.
- Corpus frequency. Frequency of each wordform is retrieved from two corpora, the Oxford English Corpus (2.5 billion tokens) and the Oxford New Monitor Corpus (9.3 billion tokens). These corpora are complementary, with different strengths of genre, domain, and regional coverage.
- Frequency in OED citations database (Incomings). Frequency is retrieved from OED's Incomings database, which contains over 3 million manually collected citations. These citations do not constitute a balanced corpus, but they draw from an extremely diverse array of sources, including film scripts, personal letters, literature, and other sources which are not widely available in digital form. Quotations in this database are primarily from the late 20th and early 21st centuries, but also include historical evidence from as early as the Middle English period.

- Timestamp of first appearance in NWPE. The timestamp of the first recorded instance of the wordform being uploaded to NWPE.
- Earliest citation. The earliest date attested for the wordform is calculated with reference to first NWPE timestamp, both corpora, and Incomings citations. This is an important datum for OED prioritization in particular, since it indicates the candidate's longevity. As a historical dictionary, OED tends not to add entries for items less than about a decade old unless they are exceptionally prominent. (Examples of neologisms that were added to the OED after a relatively short time due to their extremely high profile include **podcast**, **crowdsourcing**, and—more recently—**Brexit**.)
- Lists including the wordform. The names of all input lists on which the wordform has appeared are retrieved. This provides information about the context in which a given wordform has previously been considered, which may provide information about regional or domain associations, or be an indication of sustained usage or interest.
- Failed searches. This field aggregates the number of failed searches on the oxforddictionaries.com website for a particular item. It serves as a very rough indication of reader interest in a particular candidate.

User interface. NWPE is accessed by users via a Web application. By default, records of all wordforms are displayed, but users generally apply filters to refine their results; it is also possible to search for a particular wordform. Filters include: alphabetic range of wordform; range of frequency in corpora and Incomings; range of quantity of failed searches; date range of earliest timestamp; presence/absence in any specified dictionary; presence/absence in any specified corpus; presence/absence as a suggestion in LEMUR; presence/absence on any specified input list.

Different filter criteria are selected by the user depending on the particular use case. For instance, an editor seeking to identify high-profile neologisms for urgent inclusion might filter on the criteria “Not in ODE/ Searched more than 100 times/ Appearing on latest list of trending words from monitor corpus”; one seeking to identify gaps in coverage of South Asian English vocabulary might search “Not in OED/ Appearing in a list of South Asian keywords/ At least 50hits on NMC/ Included in LEMUR”. After the filter is applied, results may also be organized based on any of the data fields, so editors take a heuristic approach, applying different filters and criteria until the most useful results are achieved. Over time, we hope to establish parameters for which criteria yield the best results for particular types of investigation, so that we can make them available as built-in options.

Figure 2

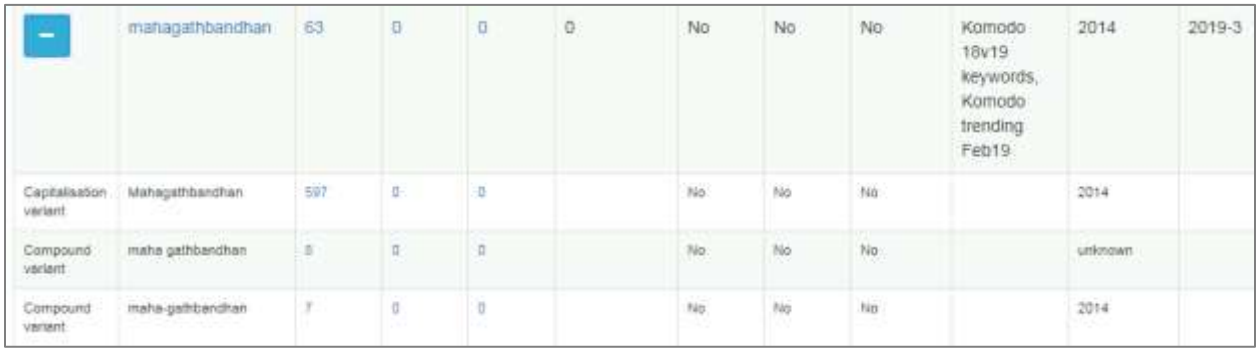


DETAIL OF RESULTS VIEW IN NWPE

Variant forms. An icon next to each wordform opens to display any potential variant forms identified in processing that occur in any of the reference inputs. This is useful for a number of reasons. First, it enables the user to see if a wordform that registers as absent from one of the reference dictionaries is in fact covered under another form. Lowercased forms of capitalized

headwords and closed forms of open or hyphenated headwords commonly register on inputs and this enables the user to disregard those wordforms as neologism candidates. However, we did not wish to filter out these near-matches altogether, because in some cases the variant may indeed represent a different lexical item. Similarly, examination of the corpus frequency of nested variants can reveal that a variant is more common than the input form, and enables the assessor to consider the aggregated evidence in making a decision about priority.

Figure 3



–	mahagathbandhan	63	0	0	0	No	No	No	Komodo 18v19 keywords, Komodo trending Feb19	2014	2019-3
Capitalisation variant	Mahagathbandhan	597	0	0		No	No	No		2014	
Compound variant	maha gathbandhan	0	0	0		No	No	No		unknown	
Compound variant	maha-gathbandhan	7	0	0		No	No	No		2014	

DETAIL OF RESULTS VIEW IN NWPE WITH VARIANTS

Validation and prioritization of wordforms. NWPE records for each wordform accept two types of modification from the users: free text comments and changes to the "Editorial Action" status. The Editorial Action field serves as both a validation record and a workflow status. The default status is "Unreviewed".

When an editor reviews at a wordform record, they first consider whether the wordform is not a valid candidate for inclusion in the dictionary. Examples of invalid words are spelling mistakes ("accomodation") and nonlexicalized proper names ("Audi"). The identification of items as invalid enables them to be excluded by filtering in future. This filtering eliminates duplication of effort when (for example) a common misspelling appears on multiple lists over a period of time. Filtering of invalid items has been especially useful for improving results from input lists of failed Web searches, since the same spelling mistakes appear consistently.

There are a number of different categories for valid suggestions, indicating different levels of priority for drafting. “Priority” candidates are recommended for immediate drafting in the dictionary. Once a valid candidate has been put forward for drafting, it is given the Editorial Action status “Actioned”, enabling it to be filtered out of results in future while also registering that it is a successful neologism.

Using NWPE to Evaluate Neologism Identification Methodologies

The Editorial Action statuses in NWPE are motivated by the practical requirements of managing workflow for multiple dictionary projects with large teams, but they also capture human judgments about neologism candidates in a way that enables comparison of different sources and methodologies to see which ones are most successful at generating high-priority candidate suggestions.

In this trial we used NWPE to compare the efficacy of two types of diachronic analysis in the Komodo corpus, a monthly monitor corpus based on online news content. We generated and uploaded to NWPE two lists. The first list contained positively trending words (according to Sketch Engine analysis) with at least 200 hits on the corpus; it contained 360 total words (the remainder of the original 1000 showed a negative trend and were therefore excluded). The second contained the top 1000 keywords with 200 hits or more on Komodo, comparing a snapshot subcorpus of the most recent four months to a reference corpus of the same four months in the previous year.

Our hypothesis was that the snapshot method, by comparing larger time-slices drawn from the same months one year apart, would be less susceptible than the trends method to topical variation. This assumption was based on the observation that high-scoring trends often relate to current events (*backstop* as used with reference to the proposed post-Brexit customs union in

Ireland) or seasonal variation (*touchdown* showing upward trends in conjunction with the beginning of the American football season). In fact, after words already covered in dictionaries were excluded, this proved not to be a significant problem for the trends list.

After the two input lists were uploaded to NWPE, those already covered in ODE/NOAD were filtered out, and the remaining items were annotated by a lexicographer. Results showed that the trending list had a slightly higher rate of both valid and invalid words, but a lower rate of words already covered in the dictionary. However, since words already in the dictionary were filtered out prior to editorial validation, the validity rate of genuine candidates on both lists was over 50%.

Table 2

	trending list	snapshot keywords list
Valid word	19.7%	15.9%
Invalid word	16.7%	11.0%
Covered under variant form	2.5%	3.4%
Already in (filtered out of results)	61.1%	69.7%

VALIDATION RESULTS OF TWO CORPUS-DERIVED CANDIDATE LISTS

There were 14 valid items that appeared on both lists, including *cume* (cumulative audience), *altcoin* (a cryptocurrency regarded as an alternative to bitcoin), *eSIM* (an integrated SIM chip for a mobile device), *tariffication* (imposition of tariffs), *stablecoin* (a cryptocurrency designed in such a way as to reduce price fluctuation), *serverless* (denoting a type of computing architecture in which servers are managed by a cloud service provider).

Items that appeared only on the trending list included *adtech*, *bispecific* (denoting a type of antibody), *cashierless* (denoting a store without human cashiers), *cryptoasset*, *enziguri* (a type of martial arts maneuver), *proptech* (property technology), *raytracing* (a rendering technique in computer graphics), and *uncrewed* (used as an alternative to 'unmanned' with reference to a vessel).

Items that appeared only on the snapshot list included *ad-supported*, *cannabis-based*, *cloud-native*, *e-scooter*, *dropbacks*, *fan-favorite*, *yellow-vest* (with reference to the gilets jaunes protests in France), *yaba* (a stimulant drug containing caffeine and methamphetamines), *ranked-choice*, and *playthrough*.

From this qualitative evaluation, it became clear that one major difference between the trends list and the snapshot list was due to the default options in Sketch Engine: the snapshot list included items with hyphens, while the trending list did not. For a like-to-like comparison, it would have been preferable to exclude hyphenated compounds from the keyword analysis. Nonetheless, the comparison demonstrated that both approaches return a high proportion of valid lexical items and are worth continuing to use as inputs in the future.

User Feedback and Future Developments

NWPE was introduced to lexicographers in a trial version in order to collect feedback to inform the next iteration of the system. Despite the limited features of the initial stage, NWPE has been widely embraced by the lexicographers on staff. They appreciate that it enables them to quickly analyze large lists of candidates that would have been too unwieldy to assess manually, assisting in a variety of specialized curation projects. The data sources which are automatically queried by NWPE have known strengths and weaknesses, which the team's lexicographers are familiar with. Aggregating data from these complementary sources makes it easier for lexicographers to

arrive at a judgment, and because all wordforms are retained in the system, they can confidently prioritize items for drafting while knowing that the less urgent items will be revisited in future.

Many ideas for enhancements have been submitted by users, and a second phase of development is planned. Users have requested additional evidence sources to enhance the data available to assist prioritization. We also hope to automate some input sources, such as trending words, alongside custom manual input lists; the trial period will be used to identify which inputs would be most valuable to automate. Improved integration with workflow systems for assigning items for drafting would increase efficiency and improve tracking of outcomes. Another desirable enhancement would be greater customization of the interface to better meet the needs of particular teams. This could be combined with preset weighted prioritization filters to improve the surfacing of the highest priority material for specific use cases, based on the heuristic solutions developed during the trial phase. Eventually, the system may also be extended to other dictionary projects, including bilinguals.

We also hope to further develop the data captured in NWPE as a collective inventory of validated and invalidated wordforms. The wordform records in NWPE could be enriched based on data captured in the system to provide information on domain and morphology even for words not (yet) selected for inclusion in any dictionary, rendering the byproduct of lexicographical decision-making a lexical resource in its own right.

References

Brookes, Ian. 2007. New Words and Corpus Frequency. *Dictionaries* 28: 142-145

Burchfield, Robert. 1973. The Treatment of Controversial Vocabulary in the Oxford English Dictionary. *Transactions of the Philological Society* 72: 1-28.

Bušta, Jan, Ondřej Herman, Miloš Jakubíček, Simon Krek, and Blaž Novak. 2017. JSI Newsfeed Corpus: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper382.pdf (accessed April 17, 2019).

Cartier, Emmanuel. 2017. Neoveille, a Web Platform for Neologism Tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* 95–98. Association for Computational Linguistics.

Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25: 447-464.

Davies, Mark. 2017. The new 4.3 billion word NOW corpus. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper250.pdf (accessed April 17, 2019).

Diamond, Graeme. 2016. Making Decisions about Inclusion and Exclusion. In *The Oxford Handbook of Lexicography*, edited by Philip Durkin, 532–545. Oxford: OUP.

Kerremans, Daphné, Susanne Stegmayr, S., and Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. In: *Current Methods in Historical Semantics,* edited by Kathryn Allan and Justyna Robinson, 59-96. Berlin: De Gruyter.

Kilgariff, Adam. 2009. Simple Maths for Keywords: http://www.kilgarriff.co.uk/Publications/2009-K-CLLiverpool-SimpleMaths.doc (accessed April 17, 2019)

Kilgariff, Adam, Jan Busta, and Pavel Rychlý. 2015. DIACRAN: a framework for diachronic analysis: https://www.sketchengine.eu/wp-content/uploads/Diacran_CL2015.pdf (accessed April 17, 2019)

New Oxford American Dictionary. 2019. In *Oxford Dictionaries*, Oxford University Press: https://premium.oxforddictionaries.com/us/english/ (accessed April 17. 2019)

O'Donovan, Ruth and Mary O'Neill. 2008. A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In *Proceedings of EURALEX 2008*, edited by Elisenda Bernal and Janet DeCesaris 571-579. Barcelon Universitat Pompeu Fabra.

Oxford Dictionary of English. 2019. In *Oxford Dictionaries*, Oxford University Press: https://premium.oxforddictionaries.com/english/ (accessed April 17. 2019)

Oxford English Dictionary Online. 2019. Oxford University Press: www.oed.com (accessed April 17. 2019)