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
Changing publication cycles
inclusion vs. prioritization

Print paradigm
Long project timeline to deliver a specific dictionary product/output

Limited number of candidates considered, mainly human-generated

What should we add?
Changing publication cycles
inclusion vs. prioritization

Digital paradigm
Short timelines to deliver maintenance updates on an ongoing basis

Vast number of potential candidates considered, human and computer-generated

What should we add first?
Changing publication cycles
inclusion vs. prioritization

**Print paradigm**
Long project timeline to deliver a specific dictionary product/output

Limited number of candidates considered, mainly human-generated

*What should we add?*

**Digital paradigm**
Short timelines to deliver maintenance updates on an ongoing basis

Vast number of potential candidates considered, human and computer-generated

*What should we add first?*

**identification and assessment**

**monitoring and prioritization**
Defining ‘neologism’
What is a ‘new word’ to a dictionary publisher?

---

**stan, v.**

Pronunciation: Brit. /stæn/, U.S. /stæn/

Frequency (in current use): •••••••

Origin: Formed within English, by conversion. Etymon: STAN n.²

Etymology: < STAN n.²

**slang (derogatory, except when self-deprecatory).**

1. **intransitive.** To be an overzealous or obsessive fan of someone, esp. a particular celebrity. Chiefly with for, over, specifying the recipient of such attention. Cf. STAN n.²


2013  @shayz27 1 July in twitter.com (O.E.D. Archive)  Anyway let me stop stanning. Or rather let me go back to secretly stanning.

2016  Billboard.com (Nexis) 15 Jan.  This isn't the first time Obama has stanned for the Compton rapper.

2. **transitive.** To be an overzealous or obsessive fan of (someone, esp. a particular celebrity).

2008  @CathrynMarie 14 Oct. in twitter.com (O.E.D. Archive)  I dnt think Bey[once] Markets to ur kind so of course ud say such, lol. not that I stan her.
Defining ‘neologism’
What is a ‘new word’ to a dictionary publisher?

bedunged, adj.

Pronunciation: Brit. /biˈdʌŋd/, U.S. /bəˈdʌŋd/, /biˈdʌŋd/

Forms:
α. ME bydyngbyd (error).
β. 16 bedungued, 16—bedunged.

Frequency (in current use): ...

Origin: Formed within English, by derivation. Etymology: be- prefix, dungen adj.

Now chiefly archaic or used self-consciously for stylistic effect.

That has been soiled with or covered in dung.

1425 Medulla Gram. (Stonyhurst) f. 33° Illitus, bydyngbyd [read bydyngyd].
1672 M. Atkins Cataphlus 50 For here they'r cramb'd up in a hole Which always lies bedung'd and foul.
1702 Winstanley's New Help to Disc. (ed. 5) 115 He sunk down dead.; the neighbours stript him, yet found no wound, but his Breeches sadly bedunged.
1785 Mem. & Adventures Flea I. vii. 199 Your property...is bedunged all over with human manure.
1868 F. J. Furnivall in Ballads from MSS (1868) I. 257 Turded farces, bedunged humbugs. [glossing Latin mimi mardati].
1929 D. Totheroh Men call Me Fool v. 44 The old peasant, in manurey smock and bedunged sabots, scratched a sparsely covered pate.
2000 Tel. & Gaz. (Worcester, Mass.) (Nexis) 29 May 13 The Royal Academy of Arts in London, which brought us the Young British Artists with their
What is prioritization?

Directing the growth of a lexical resource over time to maximize utility to users and to maintain editorial balance among the various categories of candidate items, such as general, specialist or technical, slang, and dialect terms.
Neologism monitoring

Specialized neologism trackers
• Neocrawler (Kerremans et al. 2012)
• Neoveille (Cartier 2017)

Web-based news monitor corpora
• BYU NOW
• JSI
• OUP’s NMC & Komodo
Neologism monitoring
domain balance in monitor corpora

<table>
<thead>
<tr>
<th>COCA (520 million)</th>
<th>OEC (2.5 billion)</th>
<th>JSI 2014-16 (21.3 billion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma (noun)</td>
<td>freq.</td>
<td>lemma (noun)</td>
</tr>
<tr>
<td>year</td>
<td>769,254</td>
<td>people</td>
</tr>
<tr>
<td>time</td>
<td>764,657</td>
<td>time</td>
</tr>
<tr>
<td>people</td>
<td>691,468</td>
<td>year</td>
</tr>
<tr>
<td>way</td>
<td>470,401</td>
<td>way</td>
</tr>
<tr>
<td>day</td>
<td>432,773</td>
<td>day</td>
</tr>
<tr>
<td>man</td>
<td>409,760</td>
<td>thing</td>
</tr>
<tr>
<td>thing</td>
<td>400,724</td>
<td>man</td>
</tr>
<tr>
<td>woman</td>
<td>341,422</td>
<td>life</td>
</tr>
<tr>
<td>child</td>
<td>333,849</td>
<td>part</td>
</tr>
<tr>
<td>life</td>
<td>333,085</td>
<td>work</td>
</tr>
</tbody>
</table>

TOP 10 NOUNS BY RAW FREQ. IN THREE ENGLISH CORPORAS
Practical monitoring for lexicography
Four sources for candidate identification

- user searches
- human suggestions
- corpus analysis
- targeted reading
New Words Prioritization Engine
A system to aid editors in prioritizing candidates

The use case:

- Multiple editorial teams (Oxford English Dictionary & Current English program)
- Synthesize results from multiple sources
- Exploit digital sources with a high chaff:wheat ratio
- Avoid duplication of effort
- Capture human editorial judgments to improve future performance
- Heuristically develop ranking criteria
- Unique wordforms only
NWPE system
User input examples

- Corpus keyword analysis (regional and domain)
- Diachronic corpus trends
- Morphological corpus analysis
- Lemma lists from other dictionaries
- Crowdsourced suggestions
- Reading program catchwords
- Failed searches
The NWPE pipeline
<table>
<thead>
<tr>
<th>Variants</th>
<th>Word</th>
<th>Frequencies</th>
<th>Is it in?</th>
<th>Additional details</th>
<th>Editorial actions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NMC</td>
<td>OEC</td>
<td>INCs</td>
<td>ODO failed searches</td>
</tr>
<tr>
<td>++ annexure</td>
<td>240</td>
<td>234</td>
<td>0</td>
<td>40</td>
<td>Yes</td>
</tr>
<tr>
<td>++ Banetone</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>++ Beckmann thermometer</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>++ Gentle Annie</td>
<td>23</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>++ Self-annihilating</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
</tbody>
</table>
Filtering

Sample criteria

Not in ODO + >100 failed searches + >250 NMC frequency + latest list of trending words
(high-profile neologisms for Current English dictionaries)

Not in OED + South Asian keywords + >50 OEC + earliest stamp <2005
(gaps in South Asian English for OED)
Results
Using NWPE data to evaluate source efficacy

<table>
<thead>
<tr>
<th></th>
<th>trending list</th>
<th>snapshot keywords list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid word</td>
<td>19.7%</td>
<td>15.9%</td>
</tr>
<tr>
<td>Invalid word</td>
<td>16.7%</td>
<td>11.0%</td>
</tr>
<tr>
<td>Covered under variant form</td>
<td>2.5%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Already in (filtered out of results)</td>
<td>61.1%</td>
<td>69.7%</td>
</tr>
</tbody>
</table>

Trending method: adtech, bispecific, cashierless, cryptoasset, enziguri, proptech, raytracing, uncrewed, etc.

Snapshot method: ad-supported, cannabis-based, cloud-native, e-scooter, dropbacks, fan-favorite, yellow-vest, yaba, ranked-choice, playthrough, etc.

Both methods: 14 valid items, including cume, altcoin, eSIM, tariffication, stablecoin, serverless
Results

Future development

• Additional evidence sources to enhance the data available to assist prioritization
• Automated inputs alongside custom manual input lists
• Integration with CMS
• Customization of interface
• Preset weighted prioritization filters
• Extension to other projects, including non-English
Thank you

Katherine Connor Martin
katherine.martin@oup.com

@kconnormartin
Sources
