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

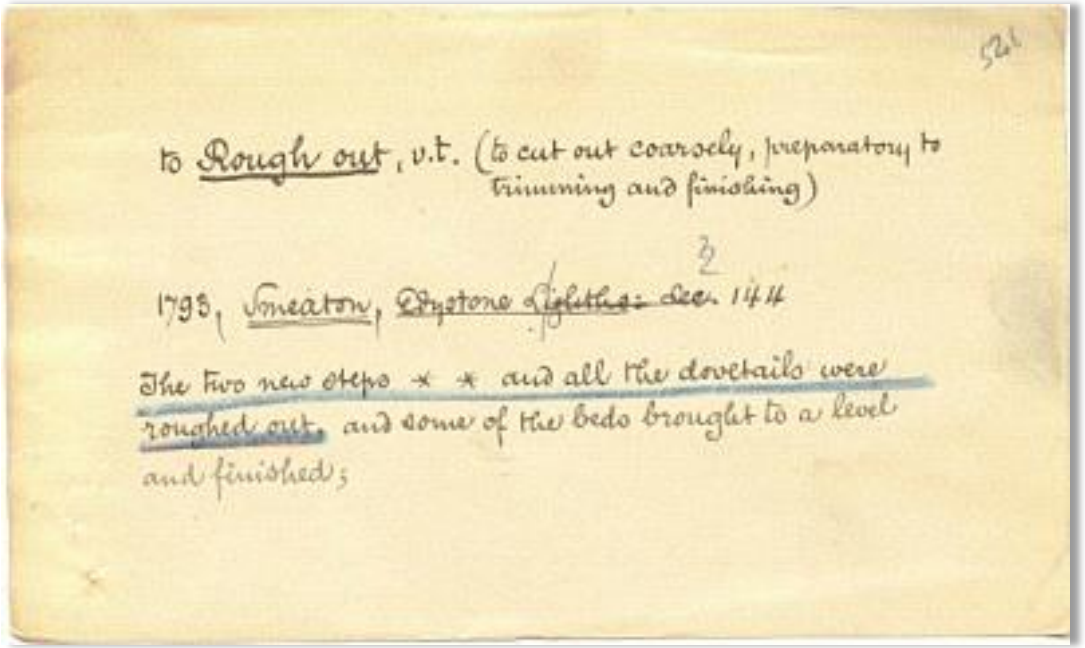
OXFORD UNIVERSITY PRESS

Print paradigm

Long project timeline to deliver a specific dictionary product/output

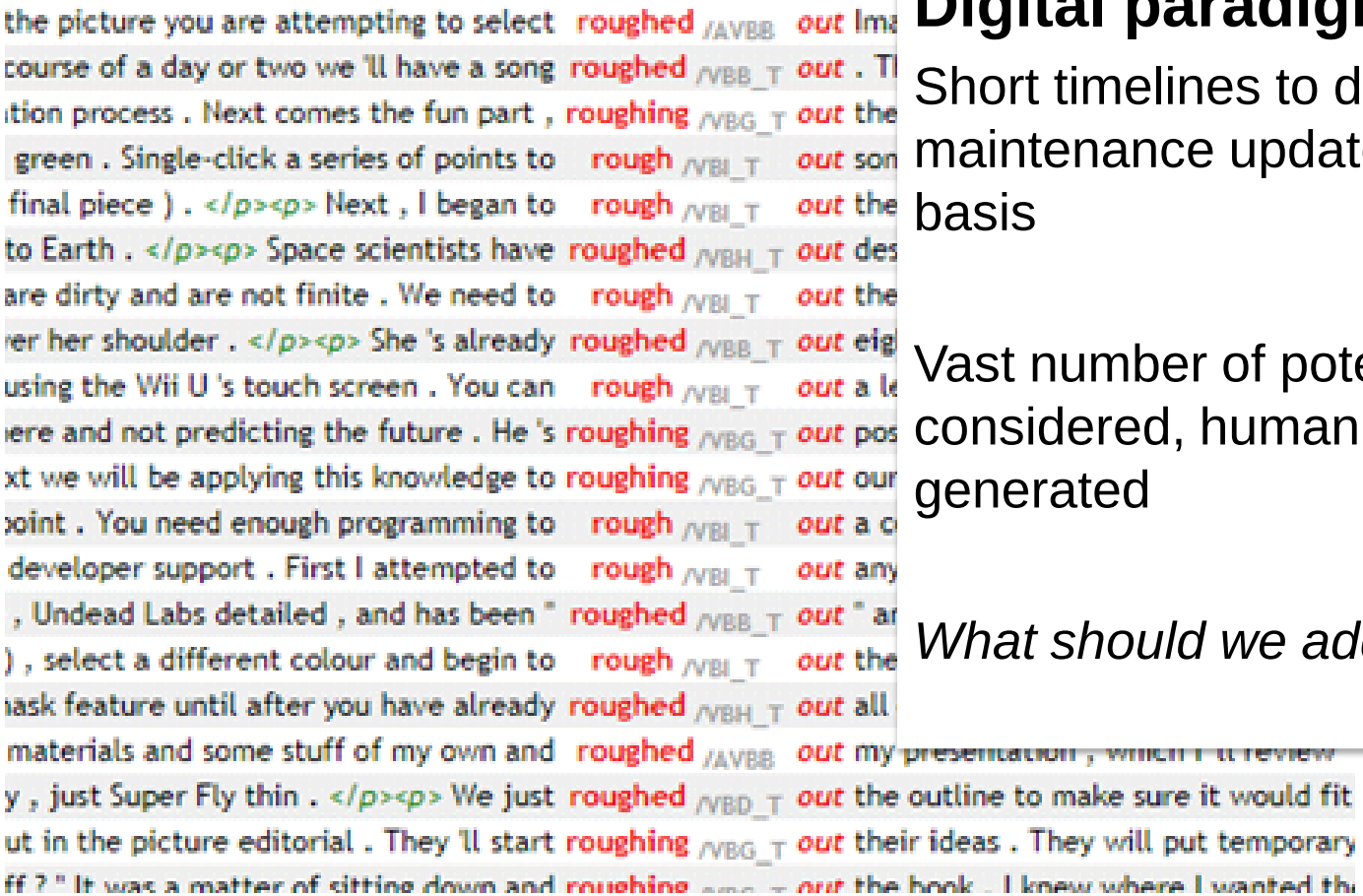
Limited number of candidates considered, mainly human-generated

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

OXFORD UNIVERSITY PRESS

Print paradigm

Long project timeline to deliver a specific dictionary product/output

Limited number of candidates considered, mainly human-generated

What should we add?

identification and assessment

Digital paradigm

Short timelines to deliver maintenance updates on an ongoing basis

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

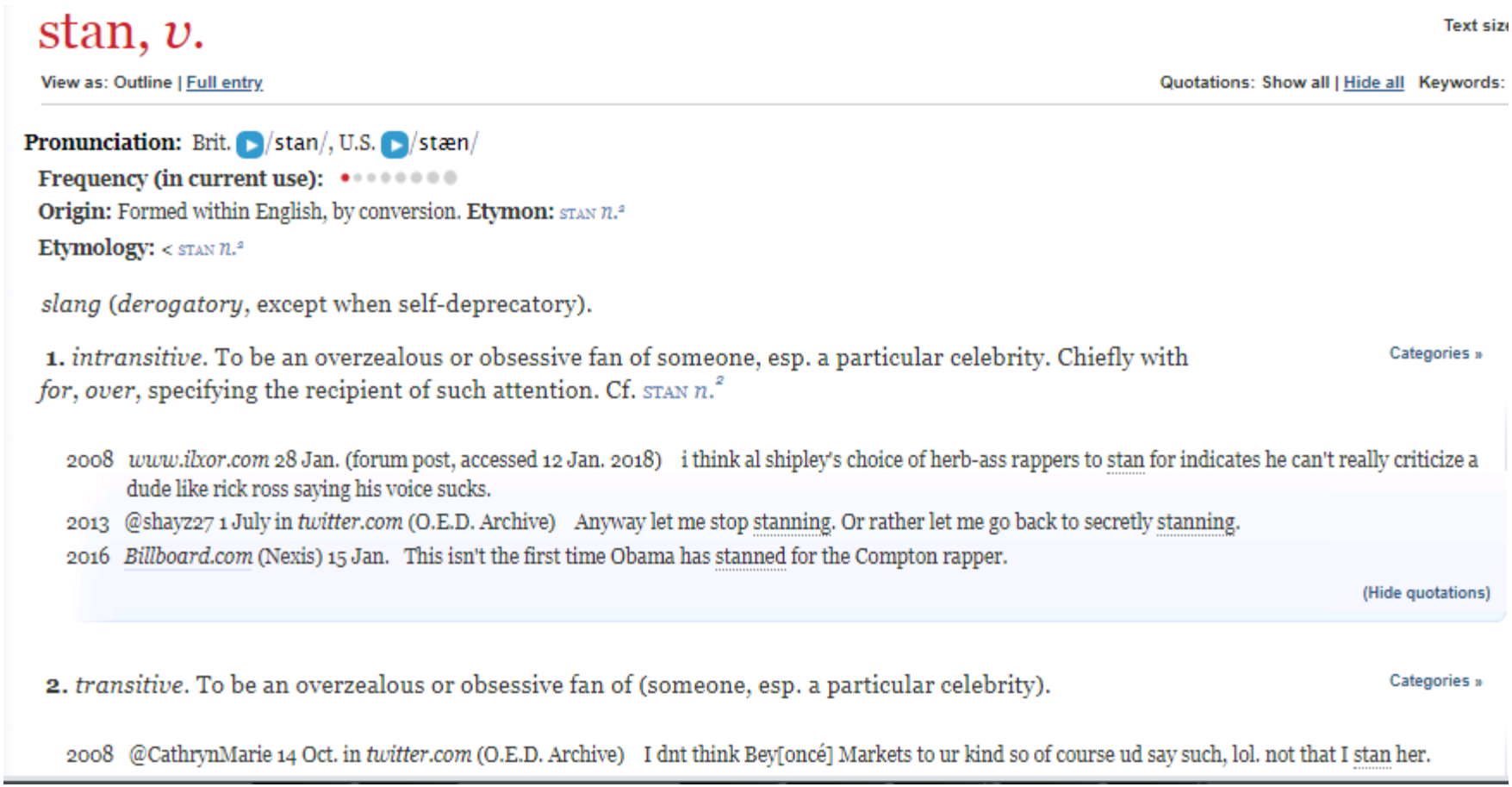
What should we add first?

monitoring and prioritization

Defining ‘neologism’

What is a ‘new word’ to a dictionary publisher?

OXFORD UNIVERSITY PRESS



stan, *v.*

Text siz

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords:

Pronunciation: Brit. /stan/, 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] Categories »

2008 *www.ilxor.com* 28 Jan. (forum post, accessed 12 Jan. 2018) i think al shipley's choice of herb-ass rappers to stan for indicates he can't really criticize a dude like rick ross saying his voice sucks.

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.

(Hide quotations)

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

2008 @CathrynMarie 14 Oct. in *twitter.com* (O.E.D. Archive) I dnt think Bey[oncé] 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?

OXFORD UNIVERSITY PRESS

---

bedunged, *adj.*

Text si

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords

Pronunciation: Brit. /bɪˈdʌŋd/, U.S. /bəˈdəŋd/, /biˈdəŋd/

Forms:

α. ME **bydyngbyd** (error).

β. 16 **bedungued**, 16– **bedunged**.

Frequency (in current use): ●●○○○○○

Origin: Formed within English, by derivation. **Etymons:** BE- *prefix*, DUNGED *adj.*

Etymology: < BE- *prefix* + DUNGED *adj.*

Compare slightly later BEDUNG *v.*

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

That has been soiled with or covered in dung.

Categories »

a1425 *Medulla Gram.* (Stonyhurst) f. 33ᵛ *Illitus*, bydyngbyd [*read* bydyngyd].

1611 R. COTGRAVE *Dict. French & Eng. Tongues* at *Bauduffle* Litter, or bedungued straw.

1672 M. ATKINS *Cataplus* 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 A3 The Royal Academy of Arts in London, which brought us the Young British Artists with their

What is prioritization?

OXFORD
UNIVERSITY PRESS

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

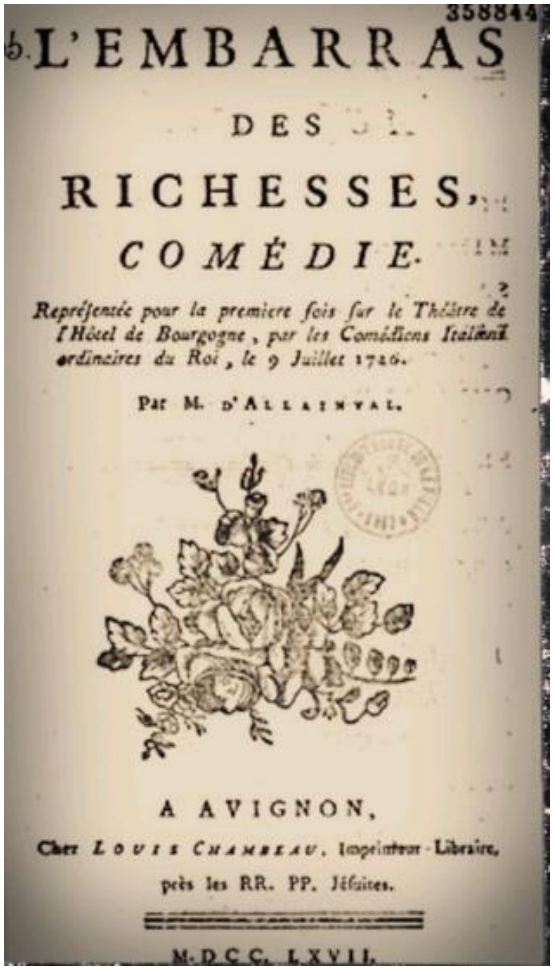
OXFORD
UNIVERSITY PRESS

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

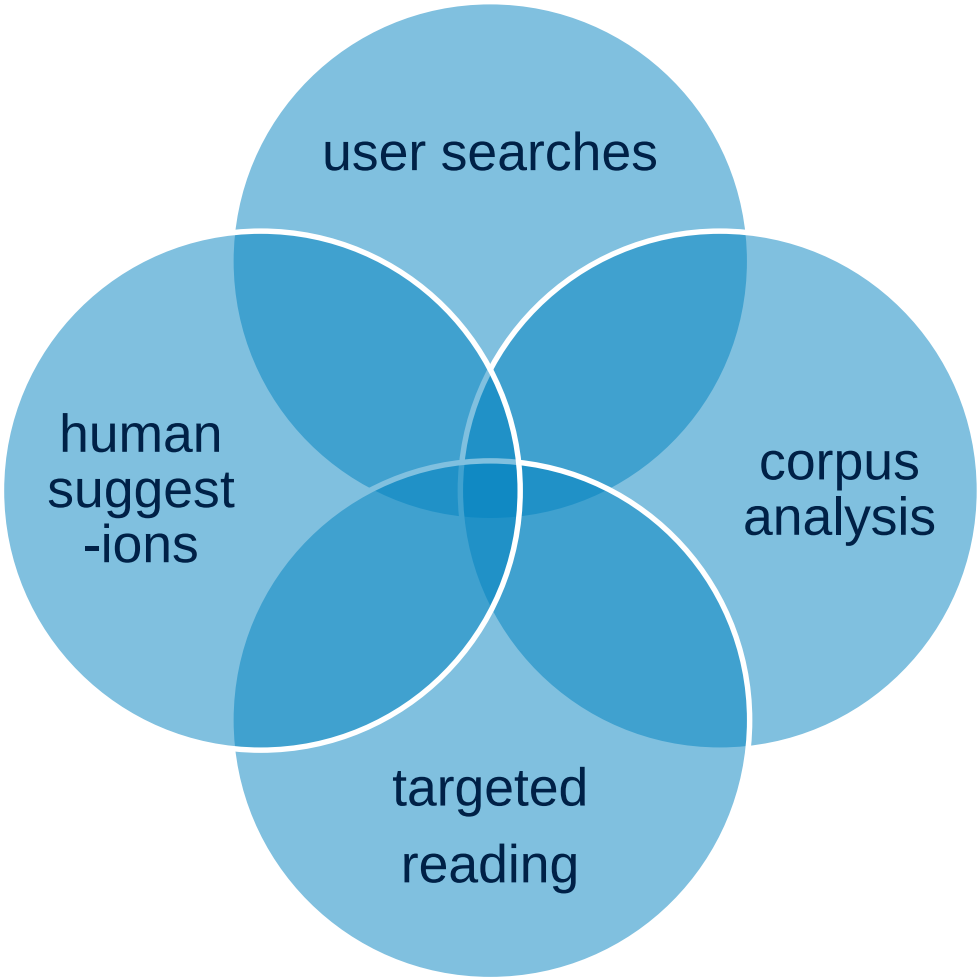
OXFORD UNIVERSITY PRESS

COCA (520 million)		OEC (2.5 billion)		JSI 2014-16 (21.3 billion)	
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 10 NOUNS BY RAW FREQ. IN THREE ENGLISH CORPORA

Practical monitoring for lexicography

Four sources for candidate identification



New Words Prioritization Engine

A system to aid editors in prioritizing candidates

OXFORD UNIVERSITY PRESS

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

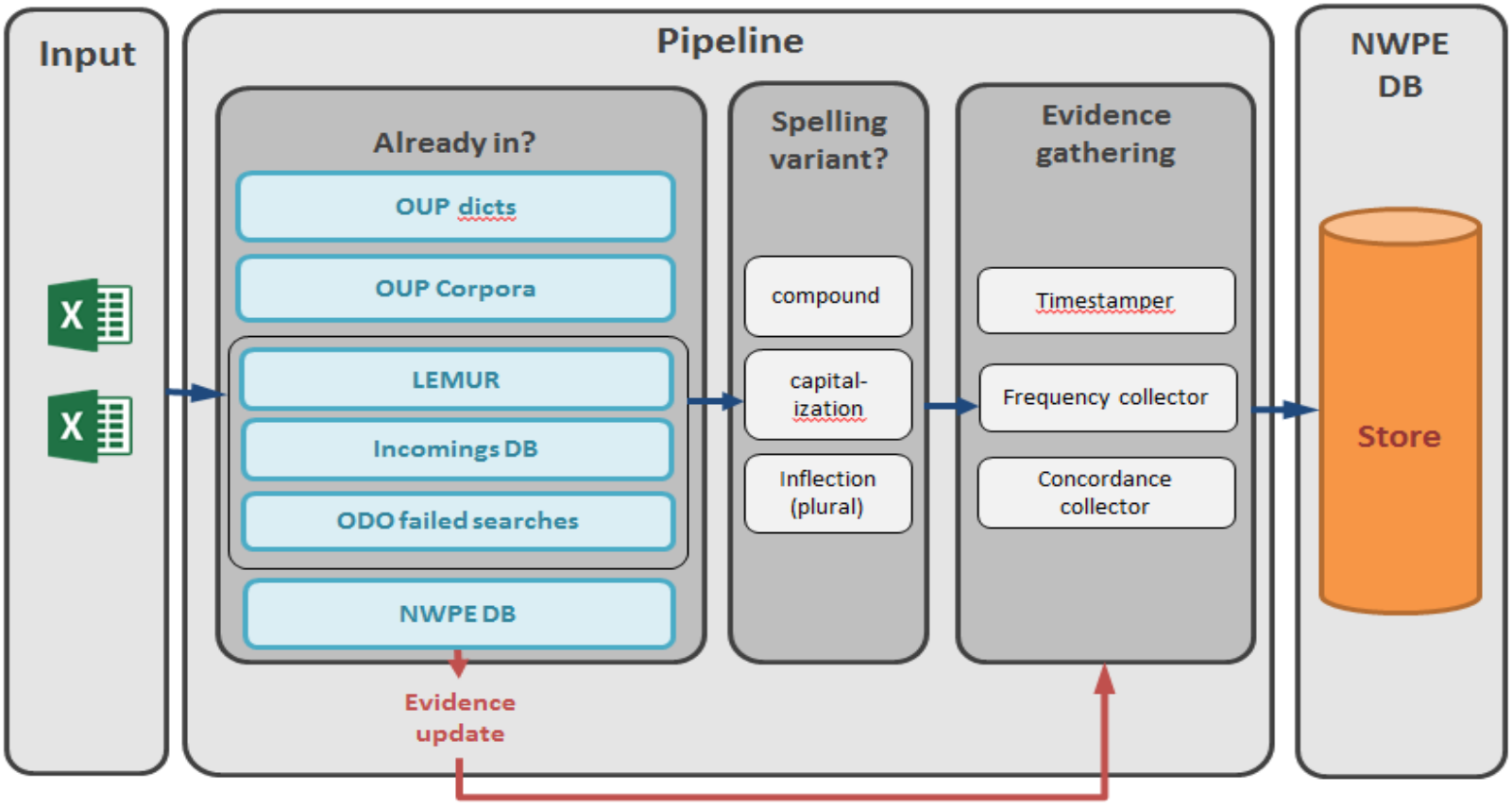
NWPE system

User input examples

OXFORD
UNIVERSITY PRESS

- 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



NWPE
Results view

		Frequencies				Is it in?			Additional details			Editorial actions	
Variants	Word	NMC	OEC	INCs	ODO failed searches	OED	ODO	Lemur	Sources	Earliest stamp	First in PE	ODO	OED
+	annexure	240	234	0	40	Yes	No	Yes	ODE failed searches 201706	2000	2017-7	actioned	already covered
+	Banneton	4	0	0	0	No	No	No	Baking words	2013	2017-7	unreviewed	unreviewed
+	Beckmann thermometer	0	1	10	0	No	No	Yes	LMR Sci suggs 160817	1909	2017-8	unreviewed	unreviewed
+	Gentle Annie	23	12	3	0	No	Yes	Yes	New Words Batch 234	1953	2017-8	unreviewed	unreviewed
+	Self-annihilating	0	0	1	0	No	No	No	selfa-selfdzz	1996	2017-7	unreviewed	unreviewed

Filtering

Sample criteria

OXFORD UNIVERSITY PRESS

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

	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%

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

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.

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.