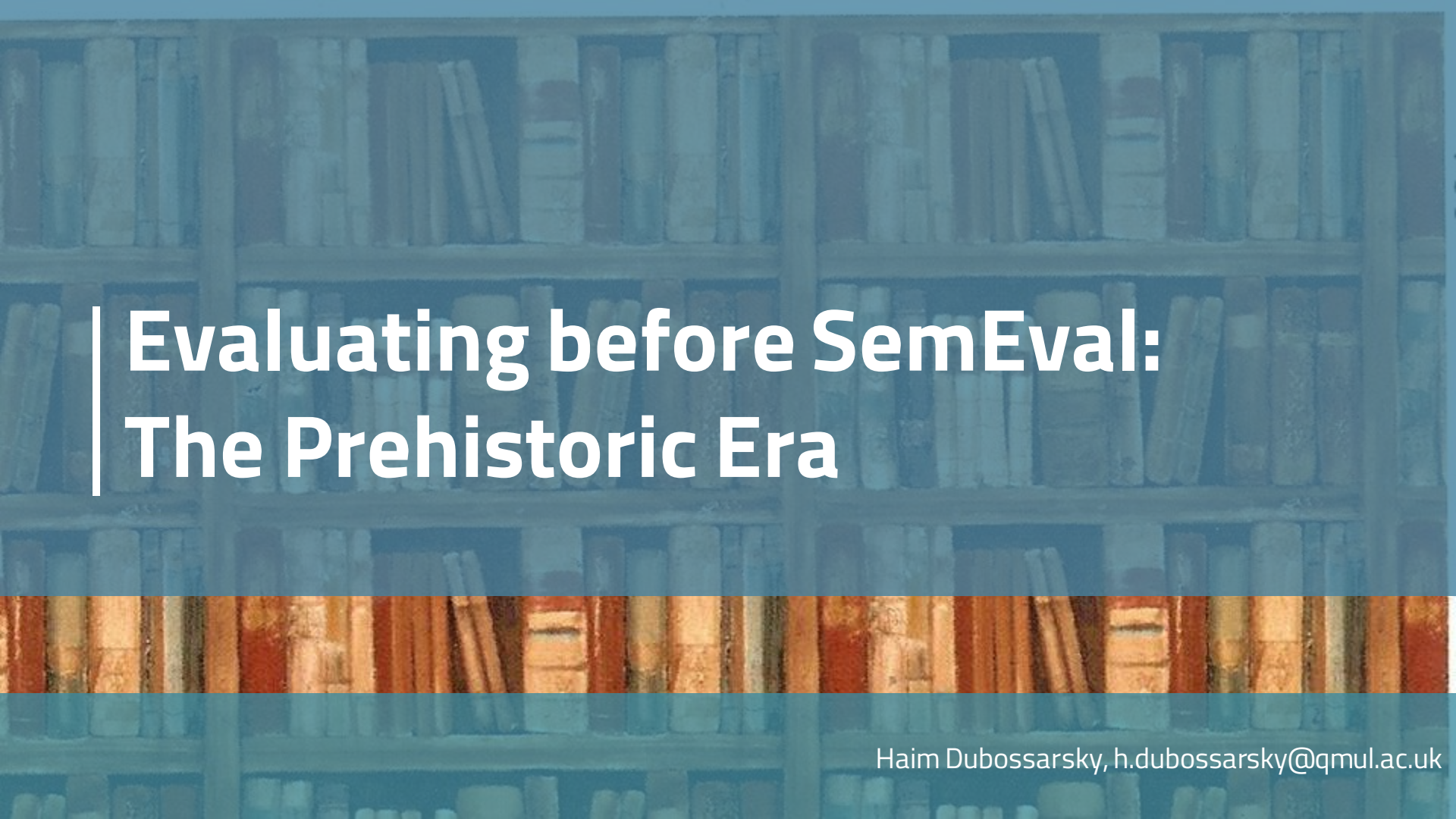
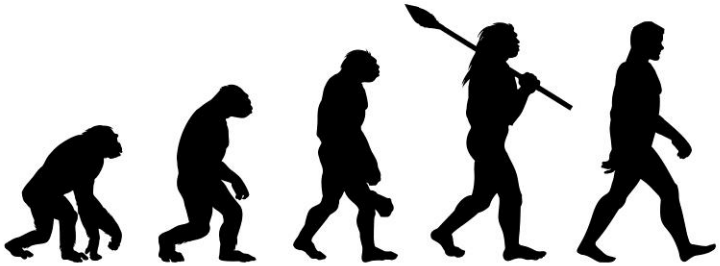
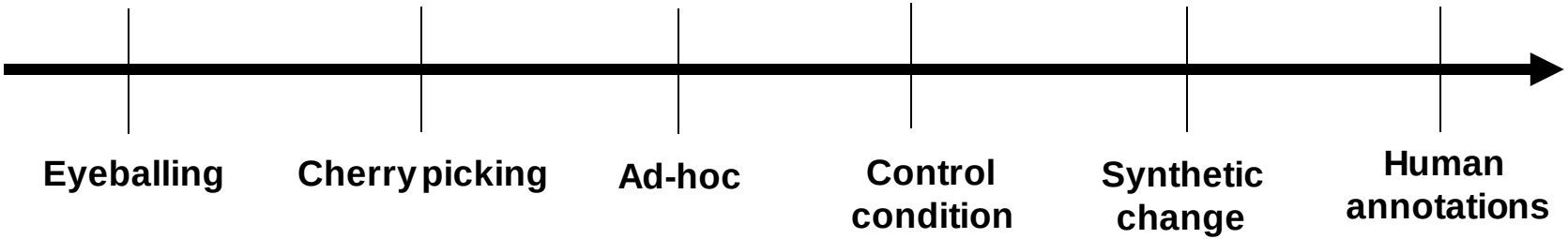


Evaluating before SemEval: The Prehistoric Era

Haim Dubossarsky, h.dubossarsky@qmul.ac.uk

A brief history of the evolution of evaluation





Qualitative: Eyeballing

Notice the similarity with PPMI

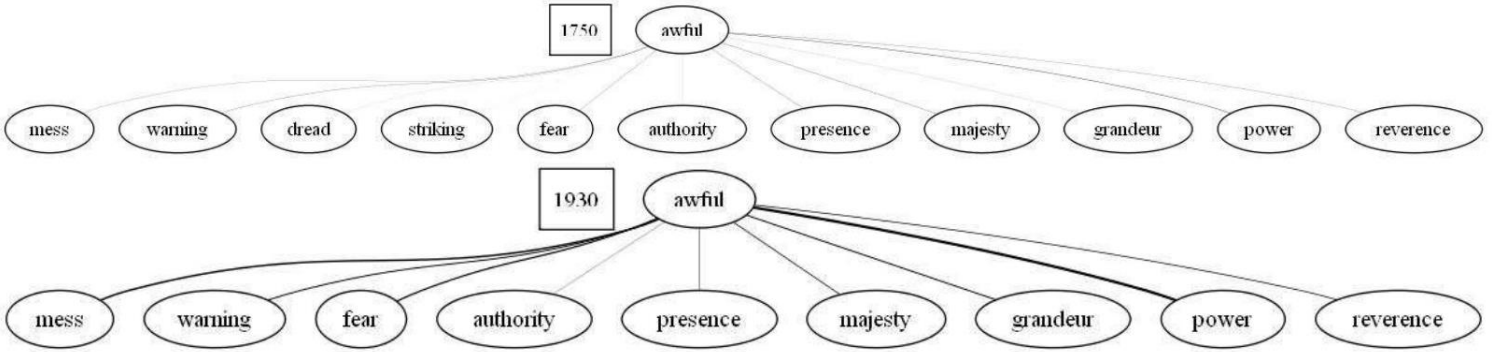


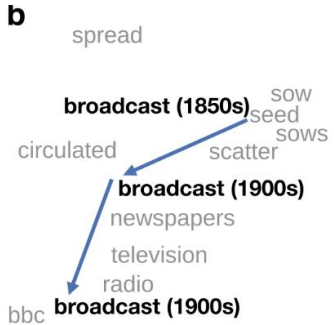
Figure 7: Words co-occurrence networks for 'awful'

From Wijaya & Yeniterzi, 2011

Qualitative: Eyeballing

Word	Neighboring Words in 1900	Neighboring Words in 2009
gay	*cheerful* *pleasant* *brilliant*	*lesbian* *bisexual* *lesbians*
cell	*closet* *dungeon* *tent*	*phone* *cordless* *cellular*
checked	*checking* *recollecting* *straightened*	*checking* *consulted* *check*
headed	*haired* *faced* *skinned*	*heading* *sprinted* *marched*
actually	*evidently* *accidentally* *already*	*really* *obviously* *nonetheless*

From Kim et al., 2014



1990s Word	1900s NN aligned with OP	1900s NN aligned with NAA	Latent Variable
wanting	need	wishing	Noise
gay	society	gay	Noise
check	give	send	Noise
starting	begin	beginning	Noise
major	general	successful	Noise
actually	believed	really	Noise
touching	touched	touching	Noise
harry	hello	john	Noise
headed	halfway	toward	Noise
romance	artists	romance	Noise
car	cab	car	Aligned
driver	stepped	driver	Aligned
eve	anniversary	eve	Aligned

Table 3: **Diachronic Semantic Change Experiment.**

From Lubin et al., 2019

Cherry picking

Word	Moving towards	Moving away	Shift start	Source
gay	homosexual, lesbian	happy, showy	ca 1920	(Kulkarni et al., 2014)
fatal	illness, lethal	fate, inevitable	<1800	(Jatowt and Duh, 2014)
awful	disgusting, mess	impressive, majestic	<1800	(Simpson et al., 1989)
nice	pleasant, lovely	refined, dainty	ca 1900	(Wijaya and Yeniterzi, 2011)
broadcast	transmit, radio	scatter, seed	ca 1920	(Jeffers and Lehiste, 1979)
monitor	display, screen	—	ca 1930	(Simpson et al., 1989)
record	tape, album	—	ca 1920	(Kulkarni et al., 2014)
guy	fellow, man	—	ca 1850	(Wijaya and Yeniterzi, 2011)
call	phone, message	—	ca 1890	(Simpson et al., 1989)

Method	Corpus	% Correct	%Sig.
PPMI	ENGALL	96.9	84.4
	COHA	100.0	88.0
SVD	ENGALL	100.0	90.6
	COHA	100.0	96.0
SGNS	ENGALL	100.0	93.8
	COHA	100.0	72.0

From Hamilton et al., 2016

Small-scale ad-hoc evaluation

group	examples	sim	freq
more frequent in 90s	users	0.29	-0.94
	sleep	0.23	-0.32
	disease	0.87	-0.3
	card	0.17	-0.1
more frequent in 60s	dealers	0.16	0.04
	coach	0.25	0.12
	energy	0.79	0.14
	cent	0.99	1.13

From Gulordava & Baroni, 2011

Word Sense Change Testset

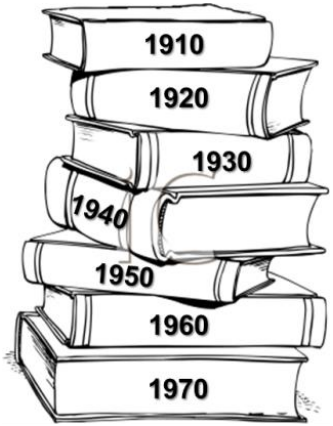
23 terms showing word sense change

From Tahmasebi & Risse, 2017

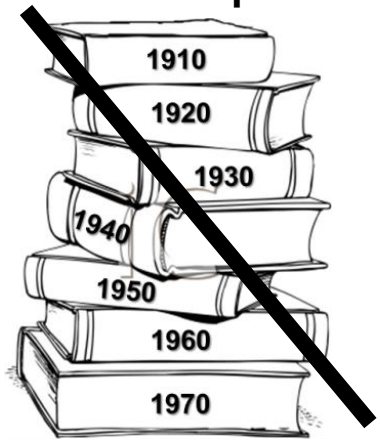
Control condition: Generation

- A control corpus resembles the genuine original historical corpus in all aspects, except what is being tested (i.e., variation in time).
- **Assumption**: any effect observed in the genuine corpus should be **lacking or reduced** in the control corpus.

Genuine corpus



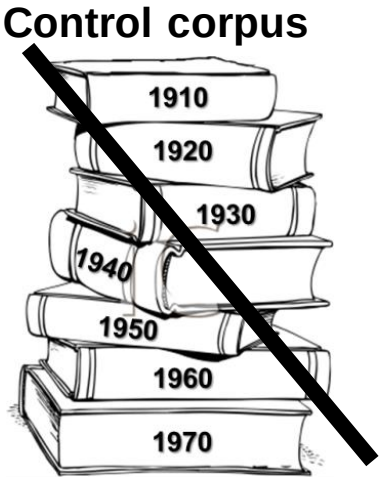
Control corpus



Control condition: Generation

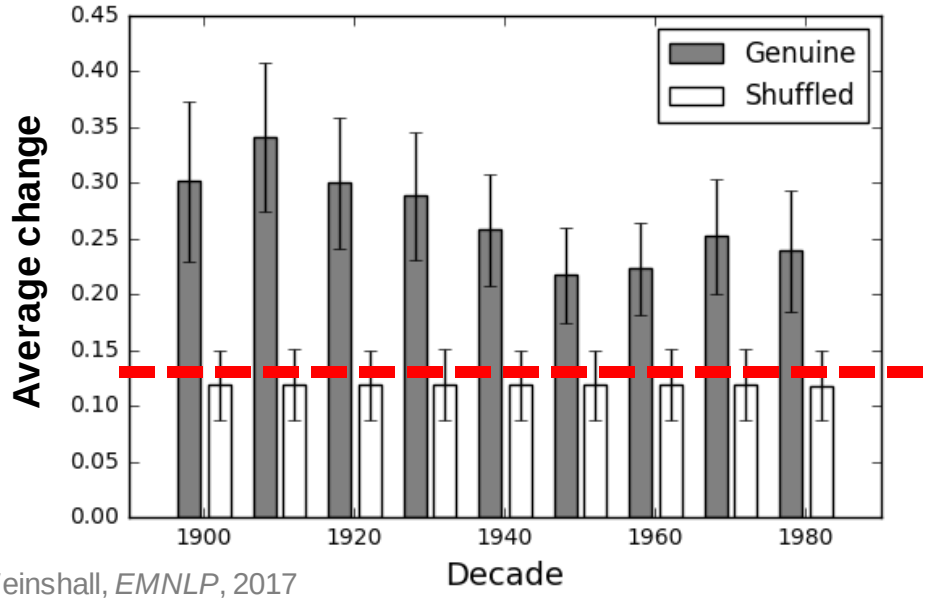
- Recipes for generating control condition:
  - Subsampling a single year's corpus (no change is assumed)
  - Shuffling between time bins of existing historical corpus (assumed changes become uniform)

- Control condition = "noise"
- Genuine condition = "noise" + "real" change
- Control condition is in fact the baseline



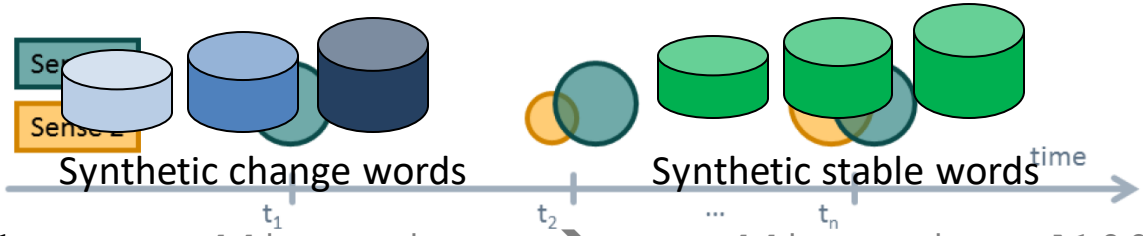
Control condition: Evaluation

The effect observed in the shuffled corpus is an artefact (model's noise)



Dubossarsky, Grossman, & Weinshall, *EMNLP*, 2017

Synthetic change: Generation



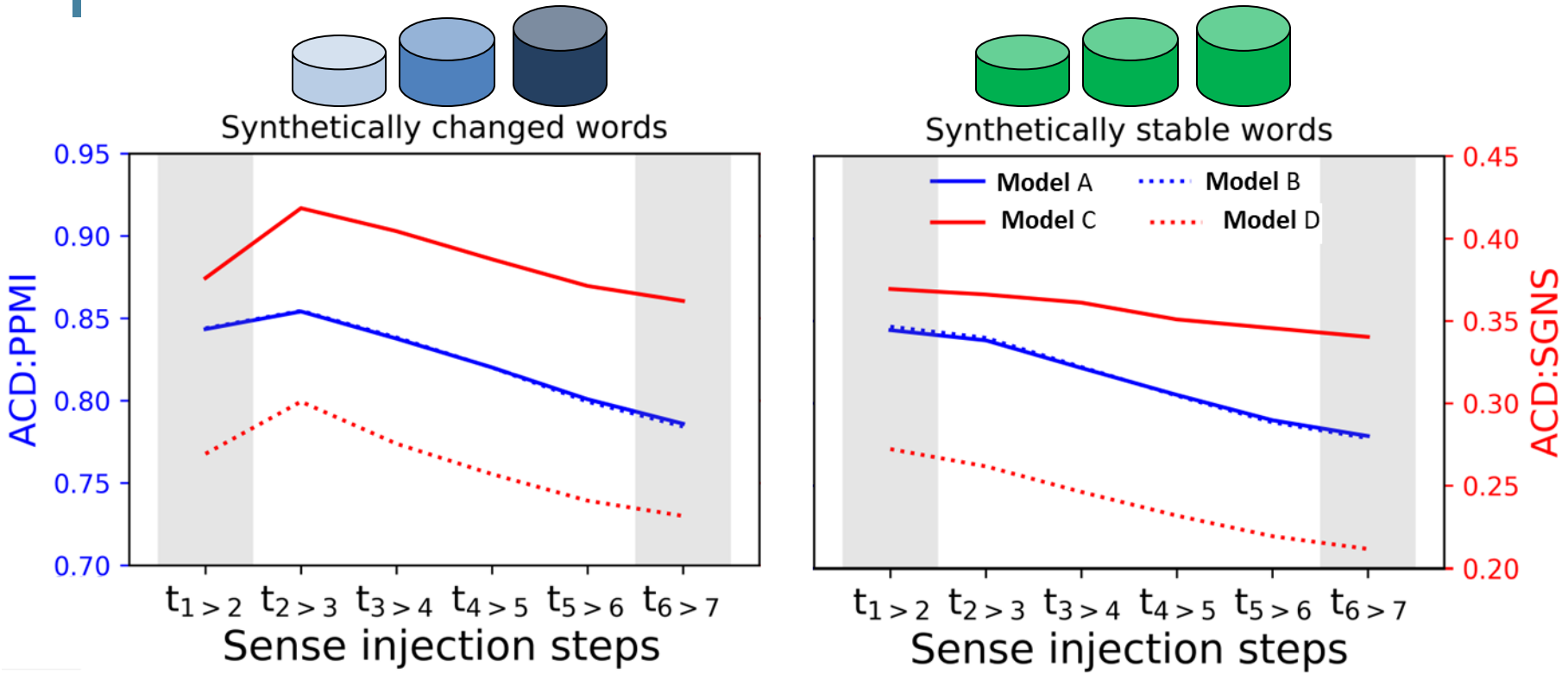
1. A wedding ring → A wedding ring [100%]
   No bracelet!
2. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [25%]
3. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [50%]

.....

4. A wedding ring → A wedding ring [100%]
   An arm bracelet → An arm ring [100%]

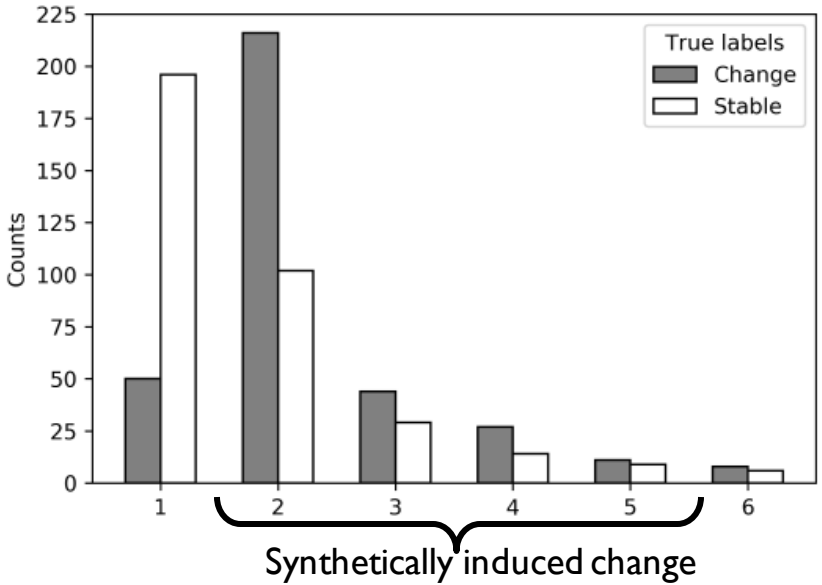
Dubossarsky et al., *EMNLP*, 2019

Synthetic change: Evaluation



Dubossarsky et al., *EMNLP*, 2019

Synthetic change: Evaluation




```
Naïve classifier

if 2=<peak_position=<5:
        semantic_change = True
else:
        semantic_change = False
```

	Model A	**Model** B	**Model** C	**Model** D
accuracy	0.65	0.66	0.59	**0.70**
F1-score	0.69	0.69	0.67	**0.74**

Dubossarsky et al., *EMNLP*, 2019

Take homes

- SemEval, while superior, is expensive and slow to develop, and limited to a specific domain, genre, or register of the language it was developed on.
- Alternatives exist, and can be useful in different research scenarios.