Data 1: SemEval2020 task 1 [1] (for evaluation of semantic change detection in our paper)
Source: https://competitions.codalab.org/competitions/20948
Description:

	Corpus 1			Corpus 2			
Language	period ($t-1$)	#tokens	avg/max/min	period (t)	#tokens	avg/max/min	#target word
English	1810-1860	25,955	701/4211/86	1960-2010	30,060	812/4,062/106	37
German	1800-1900	71,556	1,490/28,756/35	1946-1990	42,260	880/8,539/103	48
Latin	200BC-1BC	27,548	27,548/688/4,498/26	100AD-	129,568	3,239/10,362/245	40
Swedish	1790-1830	35,021	35,021/1,129/6,934/83	1895-1903	126,126	4,068/14,583/89	31

note: '#target word' means the annotated words with semantic changes (0 refers no change while 1 refers change). '#token' indicates the total token frequency of target words. 'avg/max/min' donates the average / max / min frequency for each target words.

Data2: USPTO patent data (an application example of our detection method)
Source: https://patentsview.org/download/data-download-tables
Description:

Table name	Column	Definition
g_application	patent_id	patent number
	patent_application_type	01-17 = utility application, etc.
	filing_date	date of application filing: YYYY-MM-DD
g_patent	patent_id	-
	patent_title	title of patent
g_patent_abstract	patent_id	-
	Patent_abstract	abstract text of patent
g_claims	patent_id	-
	claim_text	claim text

We concatenated the titles, abstracts, and claims of patents filed between 1960 and 2022 (outputting the dataframe as 'patent_id, content, filing_date') and further preprocessed the 'content' field (removing Greek letters, special symbols, and tokens with a length of 1), and cleaned it using a stop word list [2] specific to patent datasets. We provide the preprocessing code (./code/ patent_preprocess.py) and the relevant stopword/symbol lists in ./data/additional data for patent preprocess/.

Reference

[1] Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., Tahmasebi, N., 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection, in: Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., Shutova, E. (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation. Presented at the SemEval 2020, International Committee for Computational Linguistics, Barcelona (online), pp. 1–23.

[2] Arts, S., Hou, J., Gomez, J.C., 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy* 50, 104144.