Meanings Worth Measuring: Topical Clustering and Domain-Specific Evolution in TED Talk Semantics

November 27, 2025

Abstract

This study examines the semantic evolution of TED Talks between 2002 and 2020 using sentence embedding techniques to analyze nearly 4,000 talk transcripts. Our analysis reveals that TED Talks cluster predominantly by topic rather than time period, challenging the assumption of uniform temporal semantic drift. However, when controlling for topic, we identify divergent patterns of semantic change within domains: scientific and educational fields show increasing semantic convergence, while areas related to identity and creativity display growing diversification. Furthermore, an in-depth analysis of the concept "technology" reveals substantial semantic drift, particularly in psychological and social contexts, shifting from hardware-focused terminology toward human-centered applications. These findings demonstrate how knowledge dissemination evolves differentially across disciplines and highlight the utility of sentence embeddings for analyzing semantic change in specialized discourse communities.

Key words: Semantic change, sentence embeddings, TED Talks, diachronic analysis

1 Introduction

TED Talks have become a significant knowledge dissemination platform reaching billions of viewers and shaping public discourse. These talks offer an ideal case study of language evolution, as they explicitly aim to deliver "ideas worth spreading" in accessible formats (Gomez-Marin, 2024), employing communication strategies that likely adapt to audience expectations (Aitchison, 2011).

Recent advances in natural language processing using distributional semantics to represent words or sentences as vectors in high-dimensional space, enable quantitative analysis of semantic change (Hamilton et al., 2018; Kutuzov et al., 2018). While researchers have examined semantic evolution in general language (Hamilton et al., 2016; Frermann & Lapata, 2016), political discourse (Azarbonyad et al., 2017), and news media (Ding et al., 2023), knowledge dissemination platforms like TED remain understudied. Although (Fischer et al., 2024) observed evolving affective content in TED Talks since 2007, a comprehensive analysis of semantic change in this influential medium is lacking.

This study aims to fill this gap by applying sentence embedding techniques to analyze if and how the linguistic structure of TED Talks has evolved over time. First, we examine whether talks from different time periods exhibit distinctive semantic patterns, measuring how similarity decreases with temporal distance. Second, we investigate the importance of topic versus time. Third, we analyze within-topic temporal evolution to identify domain-specific patterns of semantic change. Finally, we conduct an in-depth analysis of how the key concept "technology" has shifted in its representation over time. Our findings contribute to both computational semantic methodology and understanding of effective science communication strategies.

2 Theoretical background and hypotheses

2.1 Semantic change in language

The study of semantic change has a long linguistic tradition (Bloomfield & Hoijer, 1965) but has been transformed by computational approaches using distributional semantics, which represents words as vectors based on their co-occurrence patterns (Harris, 1954; Firth, 1974). Firth (1974)'s principle that "you shall know a word by the company it keeps" underpins modern word embedding methods. Hamilton et al. (2018) proposed two quantitative laws of semantic change through analysis of multilingual historical corpora: the law of conformity (frequent words change meaning more slowly) and the law of innovation (polysemous words evolve more

rapidly). These findings suggested semantic change follows predictable, measurable patterns. Dubossarsky et al. (2017) subsequently challenged these laws, demonstrating that some patterns might be methodological artifacts rather than genuine semantic phenomena. Their work highlighted the need for robust control conditions and analysis techniques in semantic change research. More recent approaches have moved beyond static word embeddings to contextual embeddings, which generate different vector representations for the same word depending on its context. Martinc et al. (2020) leveraged BERT embeddings to detect diachronic semantic shift, while Card (2023) introduced a simplified approach using the most probable substitutes for masked terms. These methods capture more nuanced semantic changes than earlier approaches and demonstrate the rapid methodological evolution in this field.

2.2 Sentence-level semantic representations

While much research on semantic change focuses on individual words, sentence-level representations offer advantages for analyzing broader communication patterns. Sentence-BERT (SBERT), introduced by Reimers and Gurevych (2019), modifies the BERT architecture to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. This approach enables direct comparison of sentence meanings across different contexts or time periods. Sentence embeddings have been applied to various tasks including semantic textual similarity, information retrieval, and clustering. Baes et al. (2024) used sentence embeddings as part of a multidimensional framework for evaluating lexical semantic change. Shoemark et al. (2019) conducted a systematic comparison of semantic change detection approaches and found that using the whole time series is preferable over only comparing between the first and last time points. The application of sentence embeddings to diachronic analysis of public communication remains relatively unexplored, presenting an opportunity for novel research on how complete ideas, rather than just individual words, evolve semantically over time.

2.3 Linguistic analysis of TED Talks

TED Talks have attracted scholarly attention as a distinctive form of knowledge dissemination. Sugimoto et al. (2013) conducted one of the earliest analyses of TED Talk content, finding significant gender differences in language use. Fischer et al. (2024) conducted a data-driven analysis of affect in TED Talks, finding that talks with more positive valence and higher affective density correlated with greater popularity. They also observed that the valence of TED Talks has decreased since 2007,

while emotional density has increased in recent years. This temporal shift suggests that TED Talks are not static in their linguistic characteristics but evolve over time, possibly reflecting broader changes in communication norms. Therefore, we hypothesize:

H1: Semantic similarity between TED Talks decreases as temporal distance increases, reflecting a linguistic evolution over time.

Urooj and Alvi (2023) analyzed technical and non-technical language in TED Talks, finding that speakers strategically blend these language types to make complex ideas accessible to general audiences. They noted that TED presentations have developed into a distinct language variety characterized by deliberate simplification of technical concepts. This development suggests that linguistic changes might differ for diverse topics. To analyze this in detail, we hypothesize:

H2: TED Talks cluster more strongly by topic than by time period.

H3: Within consistent topic areas, talks show semantic evolution over time.

2.4 Methodological approaches to diachronic semantic analysis

Methodological approaches to semantic change analysis have evolved significantly over time. Early techniques based on word frequency distributions (Sagi et al., 2011) gave way to embedding-based methods that capture more nuanced semantic information. Kim et al. (2014) introduced neural word embeddings for diachronic analysis by training separate models for different time periods, while Hamilton et al. (2016) aligned embeddings across periods. These innovations enabled more precise, large-scale tracking of semantic change. Shoemark et al. (2019) demonstrated that independently trained and aligned embeddings outperform continuously trained ones for extended time periods, and emphasized analyzing complete time series rather than just endpoints. This insight revealed the non-linear nature of semantic change and the importance of examining intermediate periods. Rudolph and Blei (2018) proposed dynamic embeddings, which model embedding vectors as latent variables that drift via a Gaussian random walk over time. This approach explicitly models language change as a smooth, gradual process and avoids the need for post-hoc alignment.

Beyond temporal dimensions, Azarbonyad et al. (2017) showed that semantic

shifts occur across ideological boundaries. This finding suggests that semantic space is shaped by multiple dimensions of variation, including temporal and ideological factors. Building on this multidimensional understanding, we expect changes to differ for different concepts and examine the concept of "technology" as it is one of three key pillars of TED Talks (Gomez-Marin, 2024). Hence, we hypothesize:

H4: The key concept "technology" shows measurable semantic shifts over time.

3 Methodology

3.1 Data

This study uses TED Talk transcripts from 2002-2020 (Corral, 2025). From the full dataset of 4,005 talks, we excluded pre-2002 talks due to sparsity (fewer than 10 talks per year). We categorized the remaining talks into four periods: 2002-2005 (154 talks), 2006-2010 (712 talks), 2011-2015 (1,507 talks), and 2016-2020 (1,618 talks). Preprocessing included removing HTML tags and special characters while preserving case information, punctuation, and sentence boundaries. We segmented transcripts into sentences using NLTK's sentence tokenizer, filtering out fragments shorter than 10 characters. To investigate differences between successful and less successful talks, we normalized view counts by the number of days each talk had been online at the time of data collection:

$$\text{views per day}_i = \frac{\text{views}_i}{\text{days online}_i} \tag{1}$$

We then categorized talks into three success tiers:

$$\text{success tier}_i = \begin{cases} \text{high} & \text{if views per day}_i > Q_{0.67} \\ \text{medium} & \text{if } Q_{0.33} < \text{views per day}_i \leq Q_{0.67} \\ \text{low} & \text{if views per day}_i \leq Q_{0.33} \end{cases} \tag{2}$$

where Q_p represents the p-th quantile of the normalized view distribution. After preprocessing, we retained only talks with sufficient textual content (more than 5 sentences), resulting in our final dataset of 3,968 talks. The original dataset contained 457 unique topic tags, which we consolidated into 23 broader thematic categories. Since talks average 4.3 topics each, we assigned a "primary topic group" based on the most frequent category among a talk's tags. In cases where multiple categories appeared with equal frequency, we applied a predefined priority ordering

based on the categories' prevalence in the overall dataset. For topic grouping details, see Appendix **??** Table 5.

3.2 Semantic embedding generation

To quantify the semantic content of TED Talks, we employed Sentence-BERT (Reimers & Gurevych, 2019). Specifically, we utilized the `all-mpnet-base-v2` model, which produces 768-dimensional embeddings. We averaged the embeddings of all sentences within the talk, resulting in a single 768-dimensional vector per talk, as defined by:

$$\vec{T} = \frac{1}{n} \sum_{i=1}^{n} \vec{s_i} \tag{3}$$

where $\vec{T}$ is the talk embedding vector, n is the number of sentences in the talk, and $\vec{s_i}$ is the embedding vector of the i-th sentence. This approach allowed us to represent each talk as a point in a semantic space, with distances between points reflecting semantic differences between talks.

3.3 Temporal similarity analysis

We implemented a centroid-based temporal similarity analysis across four time periods (2002-2005, 2006-2010, 2011-2015, 2016-2020). For each period, we created a centroid vector by averaging all sentence embeddings from talks within that period. We calculated cosine similarity between these centroids for all period pairs:

$$\cos(\vec{C_{t_i}}, \vec{C_{t_j}}) = \frac{\vec{C_{t_i}} \cdot \vec{C_{t_j}}}{||\vec{C_{t_i}}|| \cdot ||\vec{C_{t_j}}||} = \frac{\sum_{i=1}^{768} C_{t_i,i} C_{t_j,i}}{\sqrt{\sum_{i=1}^{768} C_{t_i,i}^2} \cdot \sqrt{\sum_{i=1}^{768} C_{t_j,i}^2}} \tag{4}$$

We also calculated topic-specific centroids by averaging embeddings within each success tier and topic group to control for topical differences. Using linear regression, we tested whether semantic similarity decreases as temporal distance increases:

$$\text{Similarity}(t_i, t_j) = \alpha + \beta_1 \cdot |t_i - t_j| + \beta_2 \cdot \text{TopicSame} + \beta_3 \cdot \text{SuccessTier} + \epsilon \tag{5}$$

where $|t_i - t_j|$ represents temporal distance between periods, TopicSame indicates within-topic comparison, and SuccessTier controls for talk popularity.

We chose the centroid-based approach because (1) it creates robust period representations by aggregating across multiple talks, reducing the influence of outliers; (2) it generates independent observations suitable for statistical analysis; (3) it directly addresses our research question about period-level evolution; and (4) it handles our

unbalanced time periods effectively (Martinc et al., 2020). Despite these advantages, the centroid approach has limitations: it sacrifices information about within-period semantic diversity, can be sensitive to skewed topic distributions, and produces fewer data points than pairwise approaches. To address these limitations, we conduct a robustness analysis using a pairwise comparison.

3.4 Semantic clustering analysis

To test whether talks cluster more strongly by time period than by topic, we applied t-SNE (t-Distributed Stochastic Neighbor Embedding) to project the high-dimensional talk embeddings into a two-dimensional space while preserving semantic relationships:

$$\vec{T}^{\text{2D}} = \text{t-SNE}(\vec{T}) \tag{6}$$

We created separate visualizations with talks colored by time period and by topic group to visually assess clustering patterns. Further, we conducted an Analysis of Similarities (ANOSIM) test to statistically evaluate whether between-group distances are significantly larger than within-group distances for both time-based and topic-based groupings. The ANOSIM R statistic is calculated as:

$$R = \frac{r_B - r_W}{(N(N-1)/4)} \tag{7}$$

where r_B is the mean between-group rank distances, r_W is the mean within-group rank distances, and N is the total number of talks. The statistical significance of R is determined through permutation testing.

3.5 Within-topic temporal evolution analysis

For each qualifying topic (minimum 30 talks), we isolated talks assigned to that topic, controlling for the dominant topic effect observed earlier. We calculated pairwise semantic similarities between all talks within each topic using cosine similarity of their embeddings. The relationship between temporal distance and semantic similarity was modeled using linear regression:

$$\text{Similarity}_{i,j} = \beta_0 + \beta_1 \cdot \text{TemporalDistance}_{i,j} + \beta_2 \cdot \text{SuccessTier}_i + \beta_3 \cdot \text{SuccessTier}_j + \epsilon_{i,j} \tag{8}$$

where $\text{Similarity}_{i,j}$ is cosine similarity between talks i and j, $\text{TemporalDistance}_{i,j}$ is the year difference between talks, and SuccessTier_i and SuccessTier_j are success tier dummy variables.

3.6 Concept evolution analysis

To analyze how the concept of "technology" has evolved over time, we extracted all sentences containing this term, preserving their original embeddings and metadata. We calculated time-period centroids by averaging these sentence embeddings to represent how technology was discussed in each era. Semantic change was quantified using cosine similarity:

$$\text{SemanticSimilarity}(t_1, t_2) = \cos(\vec{C}_{t_1}, \vec{C}_{t_2}) \tag{9}$$

where $\vec{C}_t$ represents the centroid for period t. We controlled for confounding factors using regression:

$$\text{Similarity}_{i,j} = \beta_0 + \beta_1 \text{TemporalDistance}_{i,j} + \beta_2 \text{SameTopic}_{i,j} + \beta_3 \text{SuccessTier}_{i,j} + \epsilon \tag{10}$$

We calculated topic-specific semantic change magnitudes between earliest and latest periods:

$$\text{ChangeMagnitude}_{\text{topic}} = 1 - \cos(\vec{C}_{\text{topic},t_{\text{earliest}}}, \vec{C}_{\text{topic},t_{\text{latest}}}) \tag{11}$$

To investigate linguistic markers of this semantic evolution, we extracted the most frequent terms co-occurring with "technology" across different time periods, particularly focusing on the topic which exhibited the highest semantic change. We represent the normalized frequency of top terms across the four time periods, revealing how the contextual vocabulary surrounding technology has evolved.

4 Results

4.1 Analysis of semantic similarity over time

Figure 1 shows semantic similarity across different success tiers throughout time. Panel A shows the negative relationship between semantic similarity and temporal distance across all tiers. Panels B-E display heatmaps of cosine similarity between time period centroids for all tiers combined (B) and individual success tiers (C-E). We observe a pattern where semantic similarity decreases with time, with the earliest period (2001-2005) showing the lowest similarity to the most recent period (2016-2020). This pattern is consistent across all success tiers, though high-success talks (Panel C) exhibit a slightly steeper decline compared to medium (Panel D) and low (Panel E) success talks.

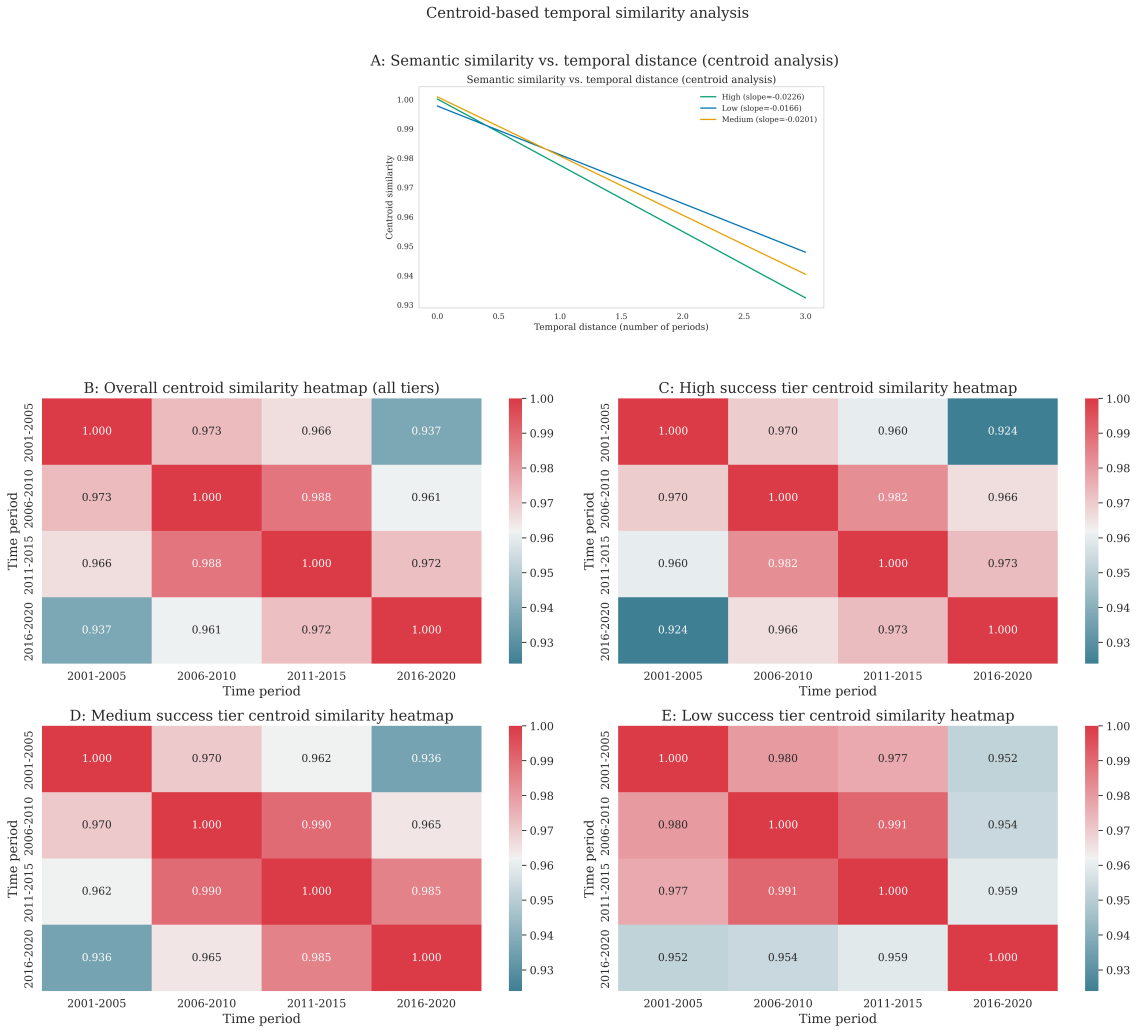

Figure 1: Centroid-based temporal similarity analysis. We observe a negative relationship between semantic similarity and temporal distance, which is the strongest for high-success talks.

Regression analysis also finds these tendencies (Model 1, Table 1), where temporal distance exhibited a positive and significant relationship with semantic similarity ($\beta = 0.070, p < 0.001$). When controlling for success of talks in Model 2, the temporal distance coefficient remained consistent and significant ($\beta = 0.070, p < 0.001$). However, when topic was introduced as a control variable in Models 3 and 4, the temporal distance coefficient became negative and non-significant ($\beta = -0.020, p = 0.256$ and $\beta = -0.020, p = 0.255$ respectively). This reversal suggests that the apparent semantic shift over time is substantially explained by thematic differences between talks rather than a genuine evolution in linguistic structure. The topic variable showed a strong negative association with similarity ($\beta = -0.189, p < 0.001$), indicating that talks addressing different topics display markedly different semantic properties regardless of when they were delivered.

	Dependent Variable: Semantic Similarity			
	Model 1	Model 2	Model 3	Model 4
Constant	0.8207***	0.8083***	0.9996***	0.9871***
	(0.006)	(0.011)	(0.025)	(0.026)
Temporal Distance	0.0697***	0.0697***	-0.0198	-0.0198
	(0.014)	(0.014)	(0.017)	(0.017)
Success Tier: Low		0.0260		0.0260
		(0.015)		(0.014)
Success Tier: Medium		0.0111		0.0115
		(0.015)		(0.013)
Same Topic			-0.1890***	-0.1889***
			(0.025)	(0.025)
Observations	298	298	298	298
R^2	0.080	0.090	0.227	0.236
Adjusted R^2	0.077	0.080	0.222	0.226

Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors in parentheses.

Table 1: Regression models of semantic similarity. Results indicate that temporal distance becomes insignificant when controlling for topic.

4.2 Semantic clustering patterns

Our clustering analysis revealed strong evidence that TED Talks cluster primarily by topic rather than by time period. Figure 2 visualizes the semantic space of TED Talks using t-SNE dimensionality reduction colored by (a) time period, (b) topic, and (c) selected major topics. While talks from different time periods are thoroughly intermixed (panel a), topic-based patterns are distinctly visible (panels b and c), with certain topics (e.g., Arts & Creativity, Health & Medicine) forming coherent regions in the semantic space. This visual assessment is corroborated by the Analysis of Similarities (ANOSIM) test results (Table 2). The ANOSIM test for topical grouping yielded a strongly positive and significant R statistic ($R = 0.601, p < 0.001$), indicating that talks within the same topic are substantially more similar to each other than to talks from different topics. In contrast, the time-based grouping produced a slightly negative R statistic ($R = -0.023, p = 1.000$), suggesting that talks from different time periods are actually more similar to each other than talks from the same period. The p-value of 1.000 for time periods provides compelling evidence that time period is not a meaningful organizing principle for TED Talk content.

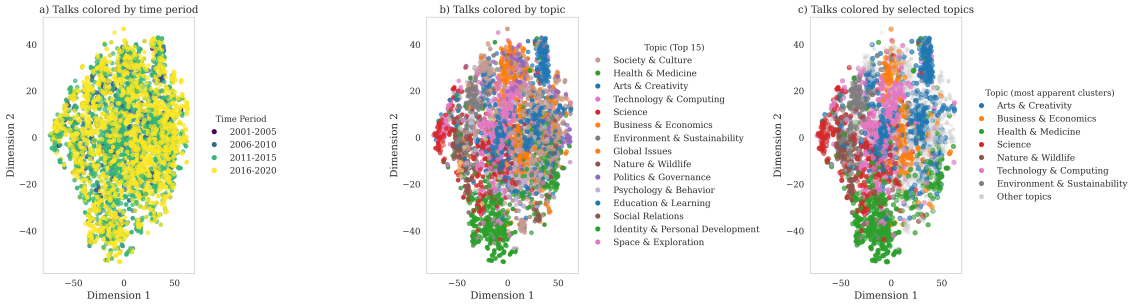

Figure 2: Talks clustered by time and topic. Visual inspection reveals that TED Talks cluster primarily by topic and not by time.

Grouping Factor	R Statistic	p-value
Topic	0.601	< 0.001
Time Period	-0.023	1.000

Table 2: ANOSIM test results for TED Talk clustering. Results show that talks within the same topic are substantially more similar to each other than to talks from different topics. Further, talks from different time periods are more similar to each other than talks from the same period.

4.3 Within topic analysis of semantic similarity over time

Our analysis of semantic evolution within individual topic areas revealed significant but divergent patterns of change over time (Table 3). Of the 16 topic areas with sufficient representation, 13 showed statistically significant temporal trends. The majority of topics (9 out of 13) demonstrated increasing semantic similarity over time. This convergence was particularly pronounced in History & Ancient Cultures ($= 0.0038, p < 0.001$). In contrast, four topics exhibited significant semantic divergence over time, with Identity & Personal Development showing the most substantial negative coefficient ($= -0.0038, p < 0.001$).

Topic	Temporal coefficient	p-value	Talks
Topics with increasing similarity over time			
History & Ancient Cultures	0.0042	0.013*	38
Psychology & Behavior	0.0038	<0.001***	117
Education & Learning	0.0028	<0.001***	113
Science	0.0024	<0.001***	312
Nature & Wildlife	0.0019	<0.001***	153
Society & Culture	0.0016	<0.001***	665
Technology & Computing	0.0012	<0.001***	502
Health & Medicine	0.0008	<0.001***	535
Environment & Sustainability	0.0006	0.015*	207
Topics with decreasing similarity over time			
Identity & Personal Development	-0.0038	<0.001***	55
Global Issues	-0.0016	<0.001***	180
Business & Economics	-0.0006	0.008**	211
Arts & Creativity	-0.0005	<0.001***	531

Note: Statistical significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 3: Regression results for semantic evolution within topics (significant effects only). Results indicate opposing trends depending on topics.

4.4 Evolution of technology concept over time

Our regression analysis of semantic change in technology-related discourse in TED Talks reveals a consistent pattern of semantic drift over time (Table 4). The base model indicates that temporal distance is significantly associated with decreased semantic similarity ($\beta = -0.0584, p < 0.001$), with each additional time period corresponding to approximately a 5.8% decrease in semantic similarity. After controlling for topic and success tier factors in Model 8, the effect of temporal distance remains highly significant ($\beta = -0.0597, p < 0.001$), suggesting robust evidence for semantic drift in technology discourse over time.

	Dependent variable: semantic similarity			
	Model 5	Model 6	Model 7	Model 8
Constant	0.9818***	0.9745***	1.0319***	1.0239***
	(0.009)	(0.009)	(0.014)	(0.019)
Temporal Distance	-0.0584***	-0.0591***	-0.0597***	-0.0597***
	(0.007)	(0.007)	(0.006)	(0.007)
Success Tier: Medium		0.0594*		0.0106
		(0.027)		(0.031)
Success Tier: Low		0.0700**		0.0212
		(0.027)		(0.031)
Same Topic			-0.0648***	-0.0568**
			(0.015)	(0.019)
Observations	163	163	163	163
R^2	0.311	0.356	0.387	0.389
Adjusted R^2	0.307	0.343	0.380	0.374

Note: $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Standard errors in parentheses.

Table 4: Regression results of semantic similarity for technology concept. Temporal distance is associated with decreasing semantic similarity across all models.

Figure 3 reveals substantial variation in how technology discourse has evolved within different domains. Psychology & Behavior displays the highest semantic change magnitude (0.462), followed by History & Ancient Cultures (0.349) and Politics & Governance (0.348). In contrast, Health & Medicine (0.043) and Technology & Computing (0.049) show remarkably little semantic change.

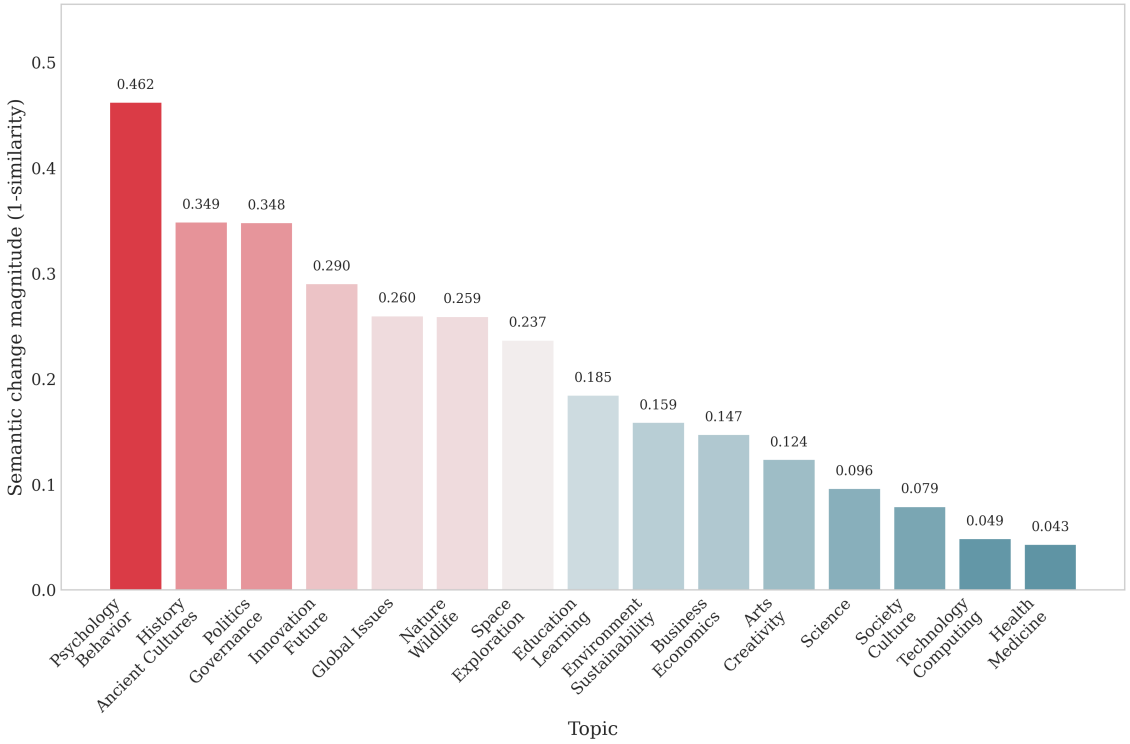

Figure 3: Semantic change magnitude by topic for the concept of technology. The change is highest for TED Talks about Psychology & Behavior and lowest for Health & Medicine.

Figure 4 depicts the evolution of terms associated with technology across four time periods in the topic Psychology & Behavior. We observe shifts in terminology over time, with earlier periods (2001-2005) emphasizing design and entertainment, while later periods show increasing association with human-centered terms. The 2011-2015 period shows strong associations with terms related to science, new developments, and brain. Most recently (2016-2020), discussions of technology have become strongly associated with terms like "help," "using," and "people".

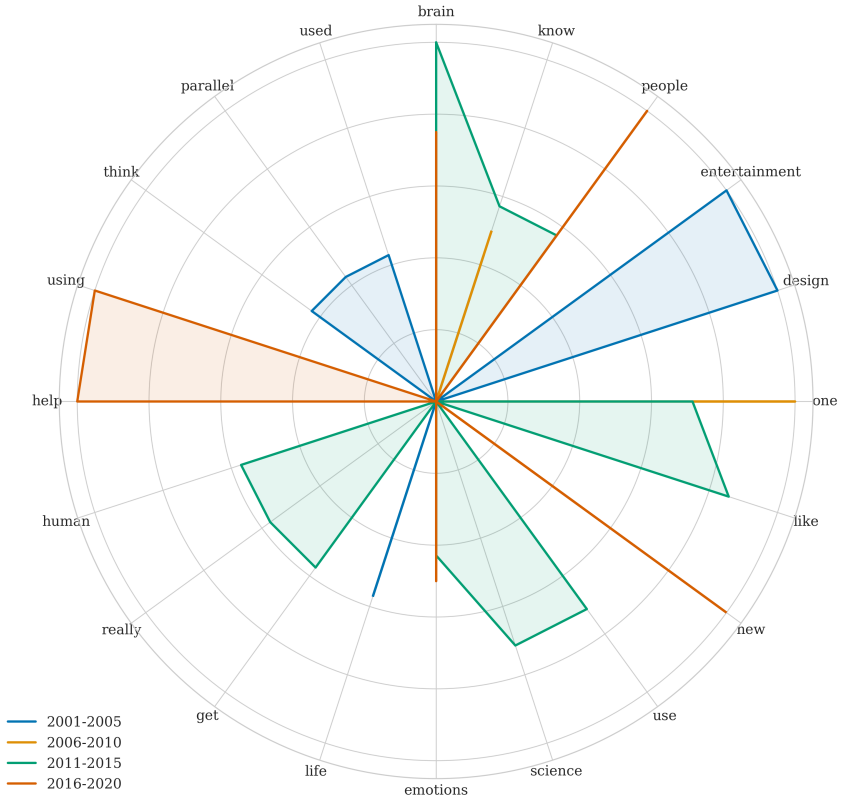

Figure 4: Evolution of terms associated with the concept "technology" in Psychology & Behavior over time. While emphasizing "design" and "entertainment" in 2001-2005, technology is more associated with "help" and "using" in 2016-2020.

5 Robustness

To validate our findings, we conducted a pairwise comparison approach analyzing talk pairs rather than aggregated time period centroids. This approach directly compares embeddings of individual talks by: (1) identifying all pairs of talks across different time periods, (2) calculating cosine similarity between each pair's embeddings, and (3) regressing similarity on temporal distance with controls for topic and success tier. Table 6 in Appendix B shows that without controls, the temporal distance coefficient is positive $(0.0023, p < 0.001)$. However, when controlling for success tier, the effect becomes negative $(-0.0002, p = 0.044)$. In the full model with topic controls, the temporal distance effect becomes non-significant $(p = 0.869)$, while same-topic pairs show substantially higher similarity (coefficient$= 0.0769, p < 0.001$). This gives further support that semantic evolution in TED Talks is primarily explained by shifts in topic distribution rather than changes in linguistic structure.

6 Discussion

6.1 Findings and contributions

This study reveals a complex interplay between topical and temporal factors in the semantic structure of TED Talks from 2002 to 2020. Our most significant finding shows that topic dominates time in determining semantic similarity between talks, which contradicts H1 and supports H2. This result aligns with Shoemark et al. (2019)'s observation that topic-specific language often overrides temporal effects in specialized discourse. As Dubossarsky et al. (2017) cautioned, apparent temporal semantic shifts may sometimes reflect changing topic distributions rather than genuine evolution in linguistic structure.

Nevertheless, our within-topic analysis uncovered meaningful temporal patterns that varied by knowledge domain, which is in line with our H3. Scientific and educational fields showed increasing semantic convergence over time, while domains centered on personal experience and creativity displayed growing diversification. This bifurcation suggests that technical fields are consolidating around standardized terminology, while humanistic areas are expanding their semantic range. This extends Hamilton et al. (2018)'s work on differing rates of semantic change by identifying trends across domains.

Our concept evolution analysis of "technology" found evidence to not reject H4 and revealed that the most substantial semantic shifts occurred where technology intersects with human behavior and social sciences, rather than in explicitly technical discussions. This supports and extends Urooj and Alvi (2023)'s observation that TED presenters strategically adapt technical language for general audiences, showing that this adaptation has evolved differently across disciplinary boundaries. The observed shift from hardware-focused terminology such as "design" to more human-centered words such as "people" and "help" in talks about Psychology & Behavior indicates a potential trend toward more practical, solution-oriented framings.

Our work challenges simplistic models of semantic evolution and suggests that different knowledge domains respond differently to similar cultural and technological pressures. Methodologically, we show the utility of sentence embeddings for tracking semantic change at multiple levels of analysis, capturing shifts in how complete ideas are expressed rather than just individual words, extending work by Frermann and Lapata (2016) and Card (2023). For practitioners in knowledge dissemination and science communication, our results suggest that effective strategies may vary by domain. The increasing standardization in scientific domains suggests benefits to consistent terminology, while the diversification in humanistic areas indicates value

in more personalized approaches. The shift toward human-centered framing reflects broader societal shifts in how we conceptualize technology's role.

6.2 Limitations and future work

Our work is not without limitations. First, our time period categorization involves somewhat arbitrary boundaries that may not align with natural inflection points in the platform's evolution. Second, while embedding distances provide a robust measure of semantic similarity, they may not capture certain rhetorical or conceptual nuances that human analysts would recognize. Third, our concept evolution analysis focuses on explicit mentions of selected terms, potentially missing more implicit discussions or related concepts.

Future work could explore whether the patterns observed in TED Talks reflect broader trends by comparing with other knowledge dissemination platforms. More detailed analysis of linguistic features beyond semantic embeddings, such as rhetorical structures or metaphor usage, could provide deeper insights into evolving communication strategies. Additionally, the divergent evolutionary trajectories across domains invite investigation into the cultural, institutional, and technological factors driving these differences. Finally, the relationship between semantic properties and talk success deserves targeted investigation.

7 Conclusion

This study has mapped the semantic landscape of TED Talks (2002-2020), revealing that thematic content dominates temporal factors in organizing this discourse space. While talks cluster primarily by topic, meaningful evolution occurs within domains. Scientific fields converge toward standardized language while humanistic domains diverge toward greater semantic diversity. The shifting representation of technology from technical capabilities toward human applications reflects broader trends in specialized knowledge communication. By applying sentence embeddings to diachronic analysis, this work advances our understanding of how semantic structures in public discourse respond to changing social contexts and audience expectations, offering both theoretical insights into semantic change and practical guidance for effective knowledge communication.

References

Aitchison, J. (2011). *Language Change: Progress or Decay?* (3rd ed.). Cambridge: Cambridge University Press.

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & Kamps, J. (2017). *Words are Malleable: Computing Semantic Shifts in Political and Media Discourse.* arXiv:1711.05603.

Baes, N., Haslam, N., & Vylomova, E. (2024). *A Multidimensional Framework for Evaluating Lexical Semantic Change with Social Science Applications.* arXiv:2406.06052.

Bloomfield, H., & Hoijer, L. (1965). *Language History from Language.* New York, New York: Holt, Rinehart and Winston.

Card, D. (2023). Substitution-based Semantic Change Detection using Contextual Embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 590–602).

Corral, M. (2025). *TED Talks Transcripts for NLP.* Retrieved from `https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset`

Ding, X., Horning, M., & Rho, E. H. (2023). Same Words, Different Meanings: Semantic Polarization in Broadcast Media Language Forecasts Polarity in Online Public Discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 161–172.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1136–1145). Copenhagen, Denmark: Association for Computational Linguistics.

Firth, J. R. (1974). *Studies In Linguistic Analysis.* Oxford: Wiley–Blackwell.

Fischer, O., Jeitziner, L. T., & Wulff, D. U. (2024). Affect in science communication: a data-driven analysis of TED Talks on YouTube. *Humanities and Social Sciences Communications*, *11*(1), 1–9.

Frermann, L., & Lapata, M. (2016). A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, *4*, 31–45.

Gomez-Marin, A. (2024). The life of "ideas worth spreading". , *386*(6718), 155.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2116–2121). Austin, Texas: Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.* arXiv:1605.09096.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Santa Fe, New Mexico, USA: Association

for Computational Linguistics.

Martinc, M., Novak, P. K., & Pollak, S. (2020). *Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift.* arXiv:1912.01072.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics.

Rudolph, M., & Blei, D. (2018). Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 1003–1011). Lyon, France: ACM Press.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing Semantic Change with Latent Semantic Analysis. In K. Allan & J. A. Robinson (Eds.), *Current Methods in Historical Semantics* (pp. 73–161). De Gruyter Mouton.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019). Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 66–76). Hong Kong, China: Association for Computational Linguistics.

Sugimoto, C. R., Thelwall, M., Larivière, V., Tsou, A., Mongeon, P., & Macaluso, B. (2013). Scientists Popularizing Science: Characteristics and Impact of TED Talk Presenters. *PLoS ONE*, *8*(4).

Urooj, F., & Alvi, U. (2023). A Linguistic Analysis of Technical and Non-Technical Language in TED talks on Science and Technology. *Journal of Corpus Linguistics*, *6*(2).

Appendix

A Topic grouping

Topic group	Individual topics
Technology & Computing	technology, computers, programming, software, AI, machine learning, algorithm, blockchain, cryptocurrency, hack, virtual reality, augmented reality, interface design, robot, robots, 3D printing, drones, code, web, Internet, data, telecom, engineering, industrial design, product design, security, surveillance, encryption, microsoft, Google, wikipedia, online video, social media, crowdsourcing, open-source, electricity, testing, gaming
Science	science, physics, chemistry, biology, astronomy, quantum physics, String theory, big bang, dark matter, universe, cosmos, Science (hard), molecular biology, microbiology, nanoscale, complexity, bionics, biotech, synthetic biology, genetics, DNA, rocket science, math, statistics, meteorology, weather, geology, paleontology, forensics, anthropology, astrobiology, biomechanics, discovery, time, CRISPR
Health & Medicine	health, medicine, medical research, public health, mental health, disease, cancer, healthcare, health care, illness, Surgery, heart health, neurology, brain, pharmaceuticals, vaccines, Vaccines, HIV, AIDS, pandemic, coronavirus, opioids, addiction, medical imaging, TEDMED, depression, disability, blindness, prosthetics, exoskeleton, bioethics, aging, pain, physiology, suicide, hearing, sight, Senses, smell, Alzheimer's, "Alzheimers", autism, Autism spectrum disorder, obesity, stigma, sleep, pregnancy, PTSD, narcotics, ebola, virus, human body, bacteria, microbes, epidemiology
Environment & Sustainability	environment, climate change, sustainability, conservation, ecology, green, biodiversity, pollution, alternative energy, solar energy, wind energy, oceans, biosphere, natural resources, nuclear energy, energy, oil, global commons, water, rivers, glacier, coral reefs, natural disaster, disaster relief, plastic, sanitation, Anthropocene, mission blue, mining, resources

Topic group	Individual topics
Society & Culture	society, culture, world cultures, diversity, inclusion, community, sociology, social change, humanity, Social Science, social media, entertainment, media, journalism, news, television, books, book, novel, literature, library, museums, consumerism, shopping, Brand, advertising, fashion, public spaces, urban, urban planning, cities, infrastructure, cooperation, collaboration, meme, poverty, inequality, language, speech, grammar, indigenous peoples
Arts & Creativity	art, music, design, creativity, film, photography, animation, theater, dance, painting, poetry, writing, literature, architecture, graphic design, typography, jazz, guitar, piano, violin, cello, vocals, composing, conducting, live music, singer, performance

Table 5: Overview of topic groupings used in the analysis.

B Robustness results

B.1 Methodology for pairwise comparison approach

To ensure our findings are not artifacts of our centroid-based methodology, we implemented a complementary pairwise comparison approach that preserves the granularity of individual talks. The methodology consists of the following steps:

1. For each success tier, we identified all pairs of talks (T_i,T_j) where T_i belongs to time period t_i and T_j belongs to time period t_j.

2. We calculated the cosine similarity between each pair of talk embeddings:

$$\cos(\vec{T_i}, \vec{T_j}) = \frac{\vec{T_i} \cdot \vec{T_j}}{||\vec{T_i}|| \cdot ||\vec{T_j}||} = \frac{\sum_{k=1}^{768} T_{i,k} T_{j,k}}{\sqrt{\sum_{k=1}^{768} T_{i,k}^2} \cdot \sqrt{\sum_{k=1}^{768} T_{j,k}^2}} \tag{12}$$

3. We constructed a regression dataset with 3,620,273 pairwise comparisons, where each observation represents the similarity between two individual talks.

4. Using this dataset, we estimated regression models with varying controls:

$$\text{Similarity}(T_i, T_j) = \alpha + \beta_1 \cdot |t_i - t_j| + \beta_2 \cdot \text{TopicSame} + \beta_3 \cdot \text{SuccessTier} + \epsilon \tag{13}$$

The pairwise approach offers several methodological advantages: By maintaining individual talk-level data rather than aggregating into centroids, we can assess whether semantic shifts persist at the individual level and avoid potential aggregation biases. Further, the substantially larger dataset (3,620,273 observations compared to 298 in the centroid approach) provides greater statistical power.

B.2 Comparison of centroid-based and pairwise results

When comparing our centroid-based and pairwise approaches, we observed notable differences in the magnitude and direction of temporal effects across different model specifications (Table 6) The centroid approach initially showed a positive temporal relationship with a larger coefficient ($= 0.070, p < 0.001$) compared to the pairwise approach's smaller positive effect ($= 0.0023, p < 0.001$). This suggests that at an aggregate level, centroids appear more similar across time than individual talks do. When controlling for success tier, the centroid approach maintained a strong positive coefficient while the pairwise approach showed a sign reversal to a small negative (statistically insignificant) coefficient ($-0.0002, p = 0.044$). Both approaches showed that topic has a strong effect on similarity, but with different implications for the temporal coefficient. In the pairwise approach, controlling for topic alone showed a positive temporal coefficient, while the full model rendered the temporal effect non-significant. The pairwise approach explained a smaller proportion of variance ($R^2 = 0.026$ in the full model) compared to the centroid approach. This reflects

the greater heterogeneity at the individual talk level that is smoothed out in the centroid aggregation.

Table 6: Regression models of semantic similarity in TED Talks

	Dependent Variable: Semantic Similarity			
	Model 9	Model 10	Model 11	Model 12
Constant	0.3885***	0.3764***	0.3804***	0.3689***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)
Temporal Distance	0.0023***	-0.0002*	0.0025***	0.00002
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Success Tier: Low		0.0285***		0.0274***
		(0.0002)		(0.0002)
Success Tier: Medium		0.0140***		0.0134***
		(0.0002)		(0.0002)
Same Topic			0.0769***	0.0762***
			(0.0003)	(0.0003)
Observations	3,620,273	3,620,273	3,620,273	3,620,273
R^2	0.000	0.005	0.022	0.026
Adjusted R^2	0.000	0.005	0.022	0.026

Note: $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Standard errors in parentheses.

B.3 Conclusion on methodological convergence

Despite the methodological differences, both the centroid-based and pairwise approaches converge on a critical insight: once topic is controlled for, temporal distance becomes non-significant or substantially reduced in importance, suggesting that apparent semantic shifts over time are primarily attributable to changes in topic distribution rather than evolution in linguistic structure. The convergence of these methodologically distinct approaches on similar conclusions strengthens the validity of our findings.