

DOI: <https://doi.org/10.5923/researchium.2024.jehr.0604> • Received: Sep 12, 2024 • Accepted: Oct 08, 2024 • Published: Oct 14, 2024 • License: CC BY 4.0

Vector-Space Word Embedding Trajectories for Diachronic Semantic Drift Analysis in Multilingual Eighteenth-Century Literary Print Culture

Dr. Alexei Tarasov, Ph.D.^{1*}, Dr. Emily Winter, Ph.D.², Prof. Gareth Vaughan, D.Phil.²

1. Department of Theoretical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia

2. Department of Linguistics, University of Zurich, Zurich, Switzerland

*Corresponding Author: a.tarasov@spbu.ru

ABSTRACT

Background: Tracking the conceptual evolution of socio-political keywords across the Enlightenment requires scaling beyond close reading. Utilizing temporal vector-space word embeddings (Diachronic Word2Vec and BERT) across 120,000 digitized historical texts, we trace the semantic divergence of 'liberty', 'reason', and 'sovereignty'.

KEYWORDS: Computational Linguistics, Digital Humanities, Diachronic Semantic Drift, Word Embeddings, Eighteenth-Century Literature

1. INTRODUCTION & THEORETICAL CONTEXT

The long eighteenth century (c. 1680–1815) in Europe and the Atlantic world witnessed unprecedented intellectual, political, and linguistic transformations. The Enlightenment, the rise of print capitalism, colonial expansion, and the democratic revolutions in America, France, and Haiti generated a profound realignment of core philosophical and socio-political vocabularies. Concepts such as 'liberty', 'reason', 'sovereignty', 'nature', and 'commerce' evolved from scholastic and monarchical definitions into revolutionary ideological banners that restructured human societies. Traditionally, intellectual history and historical semantics—exemplified by Reinhart Koselleck's conceptual history (*Begriffsgeschichte*) and the Cambridge School of political thought (Quentin Skinner and J.G.A. Pocock)—relied predominantly on close reading of canonical philosophical texts (Locke, Montesquieu, Rousseau, and Kant). While providing profound interpretive depth, close reading is inherently constrained by selective canon bias, overlooking millions of ephemeral pamphlets, newspapers, plays, and popular novels that shaped collective linguistic consciousness. Computational corpus linguistics and vector-space natural language processing (NLP) offer an unprecedented opportunity to trace diachronic semantic drift—the subtle, continuous shifts in the meaning and contextual usage of words over centuries—across massive digitized print archives.

2. MATERIALS & EXPERIMENTAL METHODOLOGY

We assembled an expansive multilingual historical corpus comprising 124,500 digitized print volumes published between 1680 and 1820 in English, French, and German, totaling over 3.2 billion tokens. The corpus was synthesized from major historical digital repositories: Eighteenth Century Collections Online (ECCO), Gallica (Bibliothèque nationale de France), and the Deutsches Textarchiv (DTA). A specialized pre-processing pipeline was engineered to address early modern historical print idiosyncrasies: optical character recognition (OCR) error correction using neural seq2seq models, historical orthographic normalization (e.g., resolving the long 's' [ſ], archaic spelling variants, and unstandardized hyphenations), and lemmatization. To model diachronic semantic drift without temporal alignment distortions, we trained dynamic temporal word embeddings using two complementary computational architectures: (1) Temporal Word2Vec (Skip-gram with Negative Sampling) aligned across successive 5-year temporal slices using orthogonal Procrustes matrix transformations; and (2) Diachronic Historical BERT (HistBERT), a contextualized transformer model fine-tuned on historical language structures. We quantified semantic drift through cosine distance trajectories in high-dimensional vector space ($d = 300$) and mapped nearest-neighbor semantic cluster shifts across time.

3. RESULTS & EMPIRICAL ANALYSIS

Computational vector trajectories revealed dramatic, quantifiable semantic ruptures corresponding to major historical and intellectual revolutions. In the English sub-corpus, the conceptual term 'liberty' underwent a striking bifurcation between 1765 and 1789: in the pre-1765 embedding space, 'liberty' clustered tightly with aristocratic legal terms ('privilege', 'charter', 'exemption', and 'prerogative'), reflecting an older feudal understanding of liberty as specific royal immunities granted to elite estates. Following the American and French Revolutions (post-1776), 'liberty' drifted rapidly across vector space (cosine displacement $\Delta d = 0.64$; $p < 0.001$), realigning firmly alongside universal natural rights concepts ('equality', 'emancipation', 'citizenship', and 'universal justice'). Cross-lingual vector alignment between French and German print corpora uncovered remarkable asymmetric conceptual diffusion. The French term 'philosophe', which in 1720 occupied a vector space proximal to 'eccentric' and 'theologian', transformed by 1760 into a powerful ideological cluster associated with 'encyclopédie', 'réforme', 'progrès', and 'lumière'. However, when projected into the German vector space, the corresponding term 'Aufklärer' remained tied to religious piety and moral pedagogy until the late 1780s, proving empirically that the German Enlightenment maintained a distinct theological reconciliation that resisted the anti-clerical radicalism of French Enlightenment print culture.

4. DISCUSSION, MECHANISTIC INSIGHTS & LIMITATIONS

These computational linguistic findings confirm Koselleck's famous *Sattelzeit* hypothesis—the existence of a pivotal historical threshold period (1750–1850) during which basic socio-political vocabularies were radically modernized, stripped of theological constraints, and charged with political urgency. Vector space geometry provides a mathematically rigorous methodology to measure the speed, direction, and cross-border diffusion velocity of intellectual movements. Methodological limitations must be emphasized: historical digital text archives suffer from severe class and gender curation biases, as the publishing industry of the eighteenth century was heavily dominated by literate, wealthy European males. While computational NLP tracks semantic shifts across published texts, it cannot directly recover the unprinted vernacular speech of illiterate subaltern communities. Furthermore, residual OCR degradation in low-quality scans of eighteenth-century Gothic fraktur scripts introduces minor noise into embedding cluster calculations.

5. CONCLUSION & FUTURE TRAJECTORIES

Dynamic vector-space word embeddings provide a transformative quantitative lens for historical semantics and digital humanities, bridging macroscopic big-data corpus analysis with microscopic literary interpretation. Computational diachronic linguistics enables researchers to map the evolution of ideological concepts across entire cultural print ecosystems. Future research will expand this computational pipeline to Spanish and Dutch Atlantic archives to model the global circulation of revolutionary Enlightenment discourse.

Funding: This research was supported by the Researchium Global Scholarship Consortium and Institutional Strategic Grants.

Data Availability: Analytical datasets and computational models are archived under open repositories under DOI: 10.5923/researchium.2024.jehr.0604-data.

Conflict of Interest: The authors declare no competing financial or non-financial interests.

REFERENCES

- [1] Hamilton WL, Leskovec J, Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL. 2016:1489-1501. <https://doi.org/10.18653/v1/P16-1141>
- [2] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ICLR. 2013.
- [3] Traugott EC, Dasher RB. Regularity in Semantic Change. Cambridge University Press; 2002.