



CASCADE

COMPUTATIONAL ANALYSIS OF SEMANTIC CHANGE ACROSS DIFFERENT ENVIRONMENTS

CASCADE Convention 1 – Short abstracts

Version: January 16, 2026



Table of Contents

CASCADE Abstracts

<i>Rachel McCarthy, University College Cork</i>	3
<i>Rasika Edirisinghe, University College Cork</i>	4
<i>Ke Shu, University of Helsinki</i>	6
<i>Yu Wu, University of Helsinki</i>	8
<i>Bách Phan-Tát, KU Leuven</i>	10
<i>Ángela María Gómez-Zuluaga, KU Leuven</i>	12
<i>Sofia Aguilar Valdez, Saarland University</i>	14
<i>Anastasiia Vestel, Saarland University</i>	15
<i>Penelope Nguyen, University of Sheffield</i>	17
<i>Maria Jimena Flores, University of Sheffield</i>	18

MECANO Abstracts

<i>Timo Zarakovitis, KU Leuven</i>	19
<i>Jonas Fischer, University of Helsinki</i>	19
<i>Luisa Ripoll-Alberola, University of Leipzig</i>	20
<i>Valeria Irene Boano, KU Leuven</i>	20



Co-funded by
the European Union



**UK Research
and Innovation**

Co-Funded by the European Union & the UKRI. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

Measuring Change in Irish Literature

Rachel McCarthy, University College Cork

“Measuring Change in Irish Literature” explores how the meanings of words have changed in Irish fiction over time. Using a dataset of 655 texts published between 1700 and 2025, this research tracks semantic and conceptual change to investigate how Irish literature reflects broader social and cultural transformations.

The first step of this analysis uses Term Frequency-Inverse Document Frequency (TF-IDF) to identify the most distinctive words across twenty-five-year time periods. By examining which terms dominate different historical periods, and by measuring the degree of similarity between each period, this research quantifies the lexical drift of Irish literature, revealing clear phases of semantic and conceptual change. The results of this initial exploration provide a baseline upon which subsequent analyses will be built. The TF-IDF results show that the vocabulary of Irish literature moves from the aristocratic and religious language of the eighteenth century towards the domestic and global vocabularies of the twenty-first.

The 1700s are characterised by both Age of Enlightenment rationalism and moral piety, with words such as “discoveries”, “philosophy”, and “deity” dominating. The 1800s maintain religious vocabularies with “monk” and “clergyman” but also expand to include words related to hierarchy and professions. Words such as “lawyer”, “heiress”, and “landlord” suggest a diversification of social identity and class representation in Irish literature. By the early 1900s, these hierarchies give way to a more modern and urban lexicon. Words such as “detective” and “murderer” mark the rise of the crime genre within Ireland and the cultural preoccupations of a newly independent state. However, the persistence of “church” and “catholic” suggests that religion remains central within Irish literature, though it is increasingly depicted in more critical contexts than previous centuries. In the 2000s, the language of Irish fiction is defined by family and everyday life. High-frequency terms such as “kids”, “dad”, and “home” show a shift to the interpersonal focus of contemporary fiction. Earlier crime-orientated narratives persist with words such as “drugs” and “investigation”, albeit recontextualised within social and domestic settings.

Cosine similarity analyses demonstrate that semantic and conceptual change in Irish literature follows a mostly gradual trajectory. Thematic tracking of five persistent concepts (family, crime, social class, religion, and rural life) shows distinct arcs. Religious vocabulary declines steadily after 1850, social-class and rural terms weaken after 1900, and crime and family themes surge in the late-1900s and early-2000s. Collectively, these changes chart a long-term transition from the hierarchical, religious worldviews of earlier centuries to the secular, globalised views of contemporary Irish fiction.

Together, these findings demonstrate that computational methods such as TF-IDF can reveal not only what Irish literature discusses, but how its language towards these concepts evolves over time. The next stage of this research will build on this foundation by applying more interpretive context-based methods, such as word embeddings, to trace how the meanings of words shift during wider social, cultural, and political movements. In doing so, this research moves from mapping broader lexical patterns to examining deeper semantic and conceptual changes in Irish literature.

Modeling Parallel Text: A Multidimensional Typology of Authorship and Transformation

Rasika Edirisinghe, University College Cork

The concept of *parallel text* has been used across literary theory, translation studies, and computational linguistics for decades, yet its meaning remains fragmented and inconsistent. This paper consolidates these divergent usages into a unified, cross-disciplinary typology that is simultaneously theoretical, computational, and visual. It reconstructs the genealogy of the term, identifies the key tensions that have shaped its evolution, and proposes a framework capable of reconciling humanistic and machine-learning perspectives within a single analytical space.

At its theoretical foundation, the study traces how relational models of textuality from Bakhtin's dialogism and Eliot's notion of tradition, through Kristeva's intertextual mosaic, Genette's transtextual taxonomy, and Riffaterre's intertextual hermeneutics established the idea that meaning arises through relations among texts (Bakhtin, 1981) (Eliot, 1975) (Kristeva, 1980) (Genette, 1997) (Riffaterre, 1981). Within these traditions, concepts such as quotation, allusion, adaptation, and transformation provide the most operationally relevant foundations for defining textual parallelism. Translation Studies later reframed parallel text as a practical resource. Early definitions by Hartmann (Hartmann, 1980), Reiss & Vermeer (Reiss, 1984), and Göpferich (Göpferich, 1999) described parallel texts as comparable originals that share genre and communicative function across languages, while corpus-based research in the 1990s, led by Mona Baker (Baker, 1995), distinguished between *parallel* (translation-based) and *comparable* (non-translation) corpora. Computational linguistics extended these ideas through alignment algorithms, probabilistic translation models, and cross-lingual embeddings (Gale & Church, 1993), (Brown, Pietra, S. A., J., & Mercer, 1993) (Schütze, 1993) transforming "parallel text" into a quantifiable signal of semantic correspondence. In the era of generative AI, large language models now produce synthetic translations, paraphrases, and adaptations, machine-generated derivatives that challenge established notions of authorship and equivalence.

To integrate these disciplinary perspectives, the paper introduces a two-axis analytical model: **Authorship (Human ↔ Machine)** and **Relation Type (Aligned ↔ Transformed)**. Their intersection yields four primary categories—**Natural, Comparative, Transformational, and Synthetic**—which together capture the full continuum from human fidelity to machine creativity. These categories are complemented by four descriptive modifiers—**Intent, Mode, Granularity, and Directionality**—that specify whether a relation is deliberate or incidental, inter- or intralingual, micro- or macro-textual, and unidirectional or co-evolved. Through this schema, each text pair can be represented as a structured data object characterized by measurable attributes such as authorship provenance, semantic alignment score, and degree of transformation.

The framework reconceptualizes *parallel text* as a dual epistemic entity, simultaneously a cognitive phenomenon of interpretive relation and a computational resource for measurable correspondence. It formalizes a shared analytical vocabulary linking literary hermeneutics, translation theory, and natural-language processing, enabling textual relations to be examined as measurable configurations within a multidimensional space. By conceptualizing parallelism as a **continuous spectrum** rather than a discrete category, the model integrates alignment, transformation, and authorship into a coherent analytical structure that accommodates both human and machine-mediated forms of textual production.

In this sense, *parallel text* becomes a bridge between interpretive and computational methodologies. The framework provides the theoretical foundation for developing more transparent and interoperable systems of text-reuse detection, semantic-similarity modeling, and visualization. It emphasizes that human and machine creativity operate within the same analytical continuum—each generating relational traces that can be modeled, compared, and interpreted as part of a unified landscape of textual correspondence.

References

- Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2), 223–243.
- Bakhtin, M. M. (1981). *The dialogic imagination*. Austin, TX: University of Texas Press.
- Brown, P. F., Pietra, D., S. A., D. P., J., V., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311.
- Eliot, T. S. (1975). *Selected prose of T. S. Eliot*. New York, NY: Harcourt Brace.
- Gale, W. A., & Church. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 75–102.
- Genette, G. (1997). *Palimpsests: Literature in the second degree*. Lincoln, NE: University of Nebraska Press.
- Göpferich, S. (1999). Parallel texts in translation. *Across Languages and Cultures*, 49–62.
- Hartmann, R. R. (1980). *Contrastive textology: Comparative discourse analysis in applied linguistics*. Heidelberg: Julius Groos.
- Kristeva, J. (1980). *Desire in language: A semiotic approach to literature and art*. New York, NY: Columbia University Press.
- Reiss, K. &. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Niemeyer.
- Riffaterre, M. (1981). *Text production*. New York, NY: Columbia University Press.
- Schütze, H. (1993). Word space. *In Advances in neural information processing systems* 5, 895–902.

From Fragment to Essay: Identifying Long-Form Reprinting in Eighteenth Century Print Culture *Ke Shu, University of Helsinki*

This study investigates reprint chains and patterns across large eighteenth-century text corpora by applying computational text reuse detection to historical books and newspapers. We leverage the extensive Eighteenth Century Collections Online corpus alongside the Burney and Nichols newspaper archives (Gale, n.d.-a; Gale, n.d.-b; Gale, n.d.-c) to explore how texts were republished across different print media. Using a previously developed text reuse dataset and methodology (Ryan et al., 2023, 25-28), we identify and contextualize segments of text that recur verbatim or near-verbatim across these corpora. Our aim is to illuminate the mechanisms of textual circulation and the dynamics of influence in Enlightenment-era print culture, where reprinting and excerpting played a central role in the dissemination of ideas.

As a pilot study, we focus on a curated subset of texts by David Hume drawn from ECCO. From the full corpus of pre-identified text reuse pairs—which include various reuse types such as reprint, citation, and other intertextual borrowings—we extract all pairs involving Hume’s works. Since our interest here lies specifically in identifying reprints of Hume’s writings, we implement a series of filtering and normalization steps. First, to mitigate fragmentation caused by OCR noise, we apply a positional offset constraint to collapse overlapping or closely adjacent reuse pieces that likely originate from a single reprint instance. Next, we trace the origin of each reuse segment to determine its earliest known source. Segments not ultimately traced back to the Hume subset are excluded, allowing us to retain only those reuse instances that represent possible reprints of Hume. These candidate reprints are then mapped to their enclosing essays—pre-annotated textual units—and manually reviewed by a domain expert. Each case is evaluated to determine whether the reuse constitutes a genuine reprint, and the relative proportion of the reused material within the essay is recorded.

In the next phase, we will conduct a series of case studies to investigate Hume-specific reprint patterns. This includes analyzing the typical length of reprinted material, identifying differences between reprint practices in ECCO books and historical newspapers, and reconstructing reprint chains over time. We will also incorporate metadata such as genre to examine how reprint behavior varies across textual contexts. Alongside this, we aim to generalize a reusable pipeline capable of supporting broader research questions: distinguishing among different categories of text reuse (e.g., reprint, citation, or other), and refining the definition of the essay as a unit of analysis—particularly in terms of how much context is necessary for a reused segment to convey sufficient meaning.

In sum, this ongoing project is expected to deliver (i) a validated workflow for isolating reprints from broader text reuse detections, including a practical offset-based defragmentation strategy and expert adjudication protocol; (ii) an annotated case-study dataset centered on Hume with labels for reprint versus other reuse categories; (iii) empirical characterizations of reprint patterns—typical passage lengths, media-specific differences between ECCO books and newspapers, and representative reprint chains; and (iv) a generalized, reproducible pipeline and guidelines for calibrating the essay-level unit, enabling researchers to adapt the approach to other authors, themes, and corpora.

References:

Ryan, Y., Mahadevan, A., & Tolonen, M. (2023, December 22). A comparative text similarity analysis of the works of Bernard Mandeville. *Digital Enlightenment Studies*, 1(1), 28–58.
<https://doi.org/10.61147/des.6>

Gale. (n.d.). *Eighteenth Century Collections Online (ECCO)*. Gale Primary Sources. Retrieved November 3, 2025, from <https://www.gale.com/primary-sources/eighteenth-century-collections-online>

Gale. (n.d.). *Seventeenth and Eighteenth Century Burney Newspapers Collection*. Gale Primary Sources. Retrieved November 3, 2025, from <https://www.gale.com/primary-sources/seventeenth-and-eighteenth-century-burney-newspapers-collection>

Gale (with the Bodleian Library). (n.d.). *Seventeenth and Eighteenth Century Nichols Newspapers Collection*. Gale Primary Sources. Retrieved November 3, 2025, from <https://gale.com/cn/product-catalog/primary-sources/seventeenth-and-eighteenth-century-nichols-newspaper-collection>

Rosson, D., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y., & Tolonen, M. (2023, April 17). Reception Reader: Exploring text reuse in early modern British publications. *Journal of Open Humanities Data*, 9, 5.
<https://doi.org/10.5334/johd.101>

Bridging Distant and Close Reading: Evaluating Semantic Search for Intellectual History in 18th-Century Books

Yu Wu, University of Helsinki

The digitization of vast historical archives, such as the Eighteenth Century Collections Online (ECCO) (Tolonen et al., 2022), combined with recent advances in Natural Language Processing (NLP), offers unprecedented opportunities for tracing the dissemination of ideas at scale. However, the inherent challenges of historical texts, including significant OCR errors, diachronic semantic change, and the conceptual complexity of philosophical language, raise critical questions about the applicability of modern semantic search tools for intellectual history research. This work directly investigates this challenge by systematically evaluating the utility of current semantic search methods, specifically based on Sentence-BERT (Reimers and Gurevych, 2019), for the core task of intellectual history: tracking the reception and transformation of complex ideas (Tolonen and Spencer, 2025).

Our work began with a methodological evaluation to understand the requirements and challenges of applying semantic search to historical texts, using the philosophy of John Locke within the ECCO corpus as a robust use case. We first performed an exploratory analysis on 20 candidate quotes sampled from a pool of 1,000 frequent reuses. By sampling and annotating their top 200 non-lexical semantic hits, we were able to map the typical distribution of results, revealing the general proportion of genuine paraphrases, meaning matches, and purely topical matches. This initial evaluation provided an empirical basis to select two quotes for a deeper, application-focused analysis, one from metaphysics, the other from philosophy of language. For these two, we collected approximately 150 validated instances of conceptual reuse for each by a comprehensive annotation.

The primary outcome of this phase is the creation of two unique, curated datasets that empirically document the semantic footprint of specific, complex philosophical arguments. We are currently enriching these datasets by systematically organizing associated ESTC metadata (e.g., publication date, author, genre, place of publication) for each instance. As proof-of-concept, this stage demonstrates that a workflow combining semantic search with efficient annotation can surpass the scale of manual retrieval to discover a significant volume of valuable semantic hits, even with a very basic model. The modern embedding models have the potential to move beyond keyword searching to identify semantically related ideas at the sentence level in noisy historical texts, which could reshape the research paradigm in intellectual history.

The next phase of this research will pivot to a dual historical analysis aimed at novel insights on Locke, which also serves as the final stage of our utility evaluation. At the macro-level, we will conduct statistical analyses of the compiled metadata to map large-scale patterns, charting the temporal and geographic distribution of each concept and identifying correlations with specific genres or authors. At the micro-level, this quantitative overview will guide targeted "close readings" of the instances themselves, allowing us to analyze the specific rhetorical functions and subtle semantic shifts in how Locke's ideas were adapted, repurposed, or contested in different contexts. Ultimately, this work aims to establish a scalable workflow that synergizes large-scale computational analysis with traditional humanistic interpretation, offering a new, evidence-based paradigm for conducting intellectual history.

References

- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Association for Computational Linguistics.
- Tolonen, M., Mäkelä, E., & Lahti, L. (2022). The Anatomy of Eighteenth Century Collections Online (ECCO). *Eighteenth-Century Studies*, 56(1), 95–123.
- Tolonen, M., & Spencer, M. G. (2025). The Reception of David Hume’s Essays in Eighteenth-Century Britain. *Hume’s Essays*, 15–35. Cambridge University Press.

SynFlow: Continuous Semantics Change Analysis via Dependency Co-occurrences

Bách Phan-Tát, KU Leuven

Modern approaches in the study of semantic change often consist of applications of vector space modelling (VSM) (Periti & Montanelli, 2024; Tahmasebi et al., 2021; Tahmasebi & Dubossarsky, 2023). However, there are many drawbacks of these methods, such as the sensitivity to corpus size (Antoniak & Mimno, 2018; Sahlgren & Lenci, 2016) and the lack of interpretability (Lenci et al., 2022). I introduce a new approach that is conceptually simpler, does not require as much data as VSM, yet is more direct in interpreting the changes in different dimensions (i.e., syntactic slots) of words' usages and meanings, along with the corresponding package, SynFlow (Phan-Tát, 2025).

The advantages of our method are:

1. **Simpler and more direct:** Given a target lemma (e.g., car) and a dependency-parsed corpus, we can extract its dependency slots (e.g., adjective modifier) and slot-fillers (e.g., red) then use Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997) to measure distributional changes of the slot-fillers of individual slots. This reveals how much a word has change, and in what dimensions (i.e., slots).
2. **Disentangled dimensions:** A conceptually related approach has been pursued by McEnery et al. (2022), who relies on surface co-occurrences. However, surface co-occurrence often suffers from accidental and/or indirect co-occurrences and the arbitrary choice of the span size (Evert, 2008), I instead adopt syntactic co-occurrence (Evert, 2008; Seretan, 2011) to separate signals.
3. **Consecutive pair-wise analysis::** Although similar functions were implemented in Sketch Engine (Word Sketch Difference, 2019), it can only work with two corpora at a time so multi-period analyses require manual aggregation whereas SynFlow can do this automatically in a single pass. Moreover, while logDice (Rychlý, 2008) is well-suited for ranking collocational salience within individual slices, it does not capture profile-level dynamics. JSD over slot-filler distributions would yield a single, interpretable change estimate for each transition with an additive decomposition into per-collocate contributions that collectively sum to the total shift.
4. **Small-corpus friendly:** Most VSMS are not usable with sparse datasets. Finetuning pretrained models often carries information from the pretraining data (Underwood et al., 2025), making the final output unreliable. To demonstrate SynFlow's ability to work well with small corpora, I will use a subset of the Royal Society Corpus (Fischer et al., 2020), with an average of 215,000 tokens per period and a total vocab size of 70,000.

In the demonstration, I will walk the audience through the different steps and functions of the workflow (in Jupyter, with minimal programming requirement), from preparing the corpus to exploring the slots distribution and analysing the slot-fillers distributional shifts. We will also demonstrate other features (e.g., slot combinations, specialization grouping) and discuss future features of SynFlow (e.g., combination with type embeddings for slot-filler clustering). I also plan to benchmark SynFlow against other approaches with SemEval 2020's task 1 (Schlechtweg et al., 2020). Audiences are also encouraged to discuss possible research questions that SynFlow could help address.

References

- Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, 6, 107–119. https://doi.org/10.1162/tacl_a_00008
- Evert, S. (2008). Corpora and collocations. In *Corpus Linguistics. An International Handbook*.
- Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study.
- Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, 56(4), 1269–1313. <https://doi.org/10.1007/s10579-021-09575-z>
- McEnery, T., Brezina, V., & Baker, H. (2022). Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, 413–444. <https://doi.org/10.1075/ijcl.18096.mce>
- Menéndez, M. L., Pardo, J. A., Pardo, L., & Pardo, M. C. (1997). The Jensen-Shannon divergence. *Journal of the Franklin Institute*, 334(2), 307–318. [https://doi.org/10.1016/S0016-0032\(96\)00063-4](https://doi.org/10.1016/S0016-0032(96)00063-4)
- Periti, F., & Montanelli, S. (2024). Lexical Semantic Change through Large Language Models: A Survey. *ACM Computing Surveys*, 56(11), 1–38. <https://doi.org/10.1145/3672393>
- Phan-Tát, B. (2025). SynFlow [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.17414457>
- Rychlý, P. (2008). A Lexicographer-Friendly Association Score.
- Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 975–980. <https://doi.org/10.18653/v1/D16-1099>
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23. <https://doi.org/10.18653/v1/2020.semeval-1.1>
- Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Springer Netherlands. <https://books.google.be/books?id=I9mO0r7HbXUC>
- Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., & Hengchen, S. (2021). Computational approaches to semantic change.
- Tahmasebi, N., & Dubossarsky, H. (2023). Computational modeling of semantic change (No. arXiv:2304.06337). arXiv. <https://doi.org/10.48550/arXiv.2304.06337>
- Underwood, T., Nelson, L. K., & Wilkens, M. (2025). Can Language Models Represent the Past without Anachronism? (No. arXiv:2505.00030). arXiv. <https://doi.org/10.48550/arXiv.2505.00030>
- Word sketch difference. (2019, May 14). <https://www.sketchengine.eu/guide/word-sketch-difference-compare-words/>

Bridging diachronic lexical semantics and conceptual history: the semantic evolution of *progress* in 18th century Britain

Ángela María Gómez-Zuluaga, KU Leuven

Diachronic lexical semanticists and intellectual historians alike have long theorised about the mechanisms of conceptual change (Geeraerts, 1997; Koselleck, 2004; Kuukkanen, 2008). Specifically, in diachronic semantics, prototype theory models lexical semantic change by rejecting strict definitional boundaries. Rather, categories have central, highly representative cases, and more peripheral ones (Geeraerts et al., 2024). Similarly, in the *Begriffsgeschichte* tradition, concepts are seen as having an invariable ‘core’ and a variable ‘margin’ (Kuukkanen, 2008). Yet despite these conceptual similarities, the relationship between the two approaches has not been systematically explored, and reflections have been conducted largely in parallel. This paper brings them into dialogue through a quantitative case study of the concept of PROGRESS in 18th-century Britain.

The concept of PROGRESS has been of extensive interest for intellectual historians and philosophers (e.g., Nisbet, 2017; Spadafora, 1990; Wagner, 2016). Our analysis focuses on how PROGRESS changed from being a field-specific notion (e.g., *progress in dramatic writing*) to having a generalised, abstract, civilisation-wide conception (e.g., *the progress of man*) within the 18th century, as suggested by Spadafora (1990). We explore the syntactic constructions and collocational patterns indicating the transition and distinction between senses, reflecting the intellectual undertones of 18th-century Britain.

Methodologically, we will combine recent advances in distributional semantics (Geeraerts et al., 2024) and well-established methods in corpus-based diachronic linguistics, from both onomasiological and semasiological perspectives. To ensure relevance to intellectual history, we use the Eighteenth Century Collections Online (ECCO), a comprehensive corpus of primary sources for the analysis of historical English discourse (Tolonen et al., 2021). With the purpose of using cleaner, machine-readable texts, we specifically use the ECCO-TCP ($\approx 2,500$ texts) transcribed by the Text Creation Partnership.

Through this analysis, we aim to identify the changing objects of PROGRESS and how the concept was referred to throughout the century. Ultimately, and based on existing readings in conceptual history and philosophical texts concerning the concept of PROGRESS, we hypothesise that, during this century, the concept underwent a shift from a field-specific sense to a generalised and abstract usage, through processes such as metaphorisation and amelioration, while retaining its semantic core.

More broadly, our study illustrates how computational and distributional methods used in diachronic lexical semantics can provide intellectual historians with large-scale descriptive evidence that broadens the scope and complements traditional close-reading methods. In doing so, we also make an initial contribution to establishing the theoretical link between diachronic semantics and conceptual history, showing how the semantic evolution of PROGRESS reflects the intellectual and cultural dynamics of 18th-century Britain.

References

- Geeraerts, D. (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press. <https://doi.org/10.1093/oso/9780198236528.001.0001>
- Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K., & Lang, M. (2024). *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford University Press. <https://doi.org/10.1093/oso/9780198890676.001.0001>.
- Koselleck, R. (2004). *Futures past: On the semantics of historical time* (K. Tribe, Trans.). Columbia University Press. (Original work published 1979)
- Kuukkanen, J.-M. (2008). Making Sense of Conceptual Change. *History and Theory*, 47(3), 351–372. <https://doi.org/10.1111/j.1468-2303.2008.00459.x>
- Nisbet, R. (2017). *History of the Idea of Progress*. Routledge.
- Spadafora, D. (1990). *The idea of progress in eighteenth-century Britain*. Yale University Press.
- Tolonen, M., Mäkelä, E., Ijaz, A., & Lahti, L. (2021). Corpus linguistics and Eighteenth Century Collections Online (ECCO). *Research in Corpus Linguistics*, 9(1), 19–34. <https://doi.org/10.32714/ricl.09.01.03>
- Wagner, P. (2016). *Progress: A reconstruction*. Polity.

Modeling changing concepts with complex networks: A case study on scientific revolutions

Sofía Aguilar Valdez, Saarland University

Language change is not random—it is driven by shifting communicative goals, social structures, and domain-specific conventions (Hamilton et al. 2016, Gries 2008, Blank 2013). While various methods exist to model change in language use, such as lexical semantic change through word meaning representations from pre-trained language models (PLMs), interpretability remains a difficult task (Periti & Montanelli 2024). This impedes domain experts' ability to qualitatively assess change and adapt models to their specific needs (Beck 2024). For instance, historians care less about new NLP benchmarks to detect when words shift in meaning, but want to understand which constellation of contextual factors—such as who is writing, where, and when—drive these changes.

We address the detection of language change in the context of scientific revolutions, where shifts in language use reflect broader epistemic transitions. Specifically, we want to model changes according to interactions of contextual factors involved in change. Modeling these interactions poses the challenges of quantifying linguistic and non-linguistic features of changing concepts and predicting which conceptual evolution mechanisms are likely to be an epistemic transition. To capture these dynamics, we propose a graph-based approach that represents scientific texts as concept networks. Our approach builds on models of knowledge development as the filling of conceptual gaps in networks (Ju et al. 2020) and the (de)formation of core/periphery structures (Kedrick et al. 2024), while acknowledging the limitations of unsupervised clustering in concept modeling (Held 2022).

We would present work in progress where we analyze interactions between key features of changing scientific concepts using complex networks. Our aim is to reveal interpretable patterns in evolving scholarship and overall provide meaningful tools for a comprehensive context-aware modeling of texts in the Digital Humanities.

References

- Beck, C. (2024). Review of Tahmasebi, Borin, Jatowt, Xu & Hengchen (2021): Computational Approaches to Semantic Change. *Journal of Historical Linguistics*, 14(2), 376–384. <https://doi.org/10.1075/jhl.22063.bec>
- Blank, A. (2013). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In A. Blank & P. Koch (Eds.), *Historical Semantics and Cognition* (pp. 61–90). De Gruyter Mouton. <https://doi.org/10.1515/9783110804195.61>
- Gries, S. T. (2008). *The identification of stages in diachronic data: Variability-based neighbor clustering*.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. <https://doi.org/10.18653/v1/P16-1141>
- Held, M. (2022). Know thy tools! Limits of popular algorithms used for topic reconstruction. *Quantitative Science Studies*, 3(4), 1054–1078. https://doi.org/10.1162/qss_a_00217
- Ju, H., Zhou, D., Blevins, A. S., Lydon-Staley, D. M., Kaplan, J., Tuma, J. R., & Bassett, D. S. (2020). *The network structure of scientific revolutions*. SocArXiv. <https://doi.org/10.31235/osf.io/tga9c>
- Kedrick, K., Levitskaya, E., & Funk, R. J. (2024). Conceptual structure and the growth of scientific knowledge. *Nature Human Behaviour*, 8(10), 1915–1923. <https://doi.org/10.1038/s41562-024-01957-x>
- Periti, F., & Montanelli, S. (2024). Lexical Semantic Change through Large Language Models: A Survey. *ACM Comput. Surv.*, 56(11), 282:1-282:38. <https://doi.org/10.1145/3672393>

Linguistic Variation across Time and Text Types: Towards Unveiling Propagandistic Strategies during the Russo-Ukrainian War

Anastasiia Vestel, Saarland University

The Russo-Ukrainian War has intensified the need to understand disinformation and its societal impacts. This PhD project investigates language variation and change in Russia's propagandistic narratives about the war, focusing on how language is manipulated to align with ideological goals. We apply computational methods to detect linguistic shifts and propaganda strategies across text types and over time. For this, we use two datasets: the Wartime Media Monitor (WarMM-2022) corpus (Alyukov et al., 2023), containing Russian news on the Russo-Ukrainian War in both state-controlled and social media; and a collection of edits in a Russian Wikipedia Fork (RWFork; Trokhymovych et al., 2025), created in June 2023 by revising the original Russian Wikipedia to comply with Russia's legislation (Cohen, 2023).

While NLP methods for propaganda detection often rely on transformer-based models that require annotated data and lack transparency, the current study aims to close this gap by using interpretable methods applied to the analysis of language variation and change. As a first step, we detect linguistic features and periods of change with the help of Kullback-Leibler Divergence (KLD; Kullback & Leibler, 1951). By employing KLD to compare Russian state and social media discourse in WarMM-2022, we identify key linguistic features that distinguish propagandistic rhetoric in war-related narratives. Our findings reveal significant lexical divergence between media types, with demobilization efforts on state media and more mobilizational rhetoric on social media, which confirms previous research on Russian propaganda in the context of this war (Alyukov et al., 2024). An example of this divergence is euphemistic framing on state media (e.g., *special military operation*) as opposed to direct terms like *war* on social media. In addition, we anticipate tracing the evolution of these linguistic tactics through diachronic analysis of WarMM-2022, highlighting their adaptation to changing political contexts.

Wikipedia presents another text type, different from media: while aiming to be a neutral and objective source of information, it can also serve as a tool for knowledge manipulation, specifically in its alternative versions such as RWFork (Trokhymovych et al., 2025). Our preliminary results from a KLD study show that divergences between the original Wikipedia and RWFork resemble those between state and social media, particularly when it comes to war- and geography-related terminology (in reference to Russia-occupied Ukrainian territories). In a way, this analysis is also diachronic, as it allows us to trace linguistic changes over time by comparing two versions of Wikipedia separated by an important historical event: the start of Russia's full-scale invasion of Ukraine.

This research contributes to the broader understanding of information manipulation in politically sensitive settings and advances methods for computational propaganda detection. In the future, we plan to extend this study by applying other methods to our datasets, such as surprisal (Shannon, 1948), which models the (un)expectedness of words in particular contexts to capture more nuanced changes in the local linguistic context, and word embeddings (Mikolov et al., 2013), which will allow us to model semantic shifts.

References

- Alyukov, M., Kunilovskaya, M., & Semenov, A. (2023). Wartime Media Monitor (WarMM-2022): A Study of Information Manipulation on Russian Social Media during the Russia-Ukraine War. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 152–161). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.latechclfl-1.17>
- Alyukov, M., Kunilovskaya, M., & Semenov, A. (2024). Confuse and Normalise: Authoritarian Propaganda in a High-Choice Media Environment and Russia's Invasion of Ukraine. In P. Goode (Ed.), *Russian propaganda today: Challenges, effectiveness, and resistance* (p. in print). University of Michigan press, University of Manchester Press.
- Cohen, N. (2023, July 12). Russian Wikipedia's Top Editor Leaves to Launch a Putin-Friendly Clone. *Bloomberg.Com*. <https://www.bloomberg.com/news/articles/2023-07-12/russian-wikipedia-editor-leaves-to-launch-a-putin-friendly-clone>
- Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), Article 1. <https://doi.org/10.1214/aoms/1177729694>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. <http://arxiv.org/abs/1301.3781>
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Trokhymovych, M., Kosovan, O., Forrester, N., Aragón, P., Saez-Trumper, D., & Baeza-Yates, R. (2025). Characterizing Knowledge Manipulation in a Russian Wikipedia Fork. *Proceedings of the International AAAI Conference on Web and Social Media*, 19, 1924–1936. <https://doi.org/10.1609/icwsm.v19i1.35910>

Contextualising the Semantic Hansard: A linguistically critical account

Penelope Nguyen, University of Sheffield

This presentation provides a linguistically critical report on the Semantic Hansard corpus, which originates from a publication called the Hansard, a comprehensive dataset containing almost all UK parliamentary speeches from 1803. The SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) project enhanced this textual collection by applying the Historical Thesaurus Semantic Tagger (HTST), which systematically annotated the corpus with semantic categories based on the Historical Thesaurus of English. This semantically tagged corpus facilitates advanced semasiological and onomasiological queries not previously available with any other corpora.

Despite the evident utility of such an extensive corpus, a systematic, linguistically informed evaluation of its construction, underlying assumptions, and practical limitations is still necessary. This study is structured around three central questions: (1) How do the inherent characteristics of the Semantic Hansard as a data source (e.g., historical reporting practices, major actors, metadata structures) shape or constrain its utility for linguistic research?, (2) What specific critiques can be directed at the HTST with regards to the Semantic Hansard?, and (3) What specific questions emerge from the preliminary explorations and analyses of the corpus?

The study employs a mixed-methods approach, centered on a two-month research secondment at the University of Glasgow, home to the SAMUELS team. Qualitative methods include direct consultation with project developers, a thorough review of internal project documentation, and a hands-on, user-centered evaluation to identify practical limitations. These approaches are complemented by quantitative corpus linguistic techniques. Notably, keyness analysis is used as a bottom-up method to reveal the internal composition of the corpus and to identify any unreported or unnoticed inconsistencies in the dataset or its semantic tags.

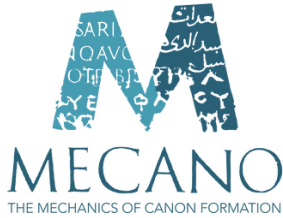
This research will culminate in a robust, linguistically critical account of the Semantic Hansard. The findings will offer valuable insights for any researcher utilizing this dataset, enabling a more informed and nuanced exploitation of the data. By articulating its inherent limitations and biases and teasing out the possibilities for exploiting the Semantic Hansard as a corpus, this study empowers researchers to formulate more realistic research goals, derive more meaningful interpretations, and make more reasonable methodological compromises.

The Historical Thesaurus of Migration: an exploration of context-driven approaches and semantic change

Maria Jimena Flores, University of Sheffield

While the phenomenon of human migration has been researched from a historical and sociopolitical standpoint, there is still a lot to study in regard to how language change reflects the shift in perception towards migration. Corpus studies, particularly diachronic corpus studies, provide the framework and the tools for the analysis of semantic change by allowing researchers to gather data on the frequency of these words over a certain period of time. In this case, starting with the colonial period in Mexico in the 16th century, progressing to the 19th century Irish transatlantic migrations to Canada and ending in the 20th century with the US-Mexico *Bracero Program*. This selection facilitates a diachronic perspective while also allowing for the diversification of the corpora through the exploration of texts in both Spanish and English. The discourses derived from each context also supply insightful data for dissecting the relationship between migration, policy and colonialism through the lens of semantic change. The result is a research project that aims to understand how the lexicon associated with migration (e.g. alien, settler or citizen) changed across time and languages by analysing the semantics of the historical contexts and identifying lexical items that are unique to each period. This is achieved through a combination of methods that capture broad semantic changes and account for data imbalances while also being context-sensitive and maintaining the interpretability of extra-linguistic factors. Early results are both promising and intriguing, with the development of a word classification system in the form of a taxonomy. The taxonomy is based on the principles of Lexical Field Analysis and the Historical Thesaurus of English to sort words into general categories related to migration that reflect the evolution of the lexicon. It was designed as a preliminary test of the lexical fields by examining the classification of words based on common attributes such as place or origin, social status or mobility. The taxonomy is also meant to be a tool for the visualization of broad semantic shifts. Its usefulness lies in the exploration of patterns and commonalities across the different time periods and languages, which also allows for the identification of outliers that are more context dependent.

Keywords: migration, semantic change, corpus studies, diachronic analysis, historical thesaurus



MECANO – Short abstracts

The philosophical canon and the art of (mis)quoting Plato and Aristotle in the Commentaria in Aristotelem Graeca

Timo Zarakovitis, KU Leuven

It is widely assumed that, at least from Iamblichus (mid 3rd–mid 4th century AD) onwards, in the Platonic schools of late antiquity philosophical education was organised around a canon of philosophical texts with works by Aristotle and Plato at the core. It is unclear what role this canon played in actual school practice and what limits it imposed on the circulation and knowledge of non-canonized Platonic and Aristotelian works. This project's goal is to complement the ancient Platonists' self-understanding of their philosophical canon with data generated by the study of citations and quotations of Plato and Aristotle in the Greek commentaries on Aristotle (most of which were written by Platonists). The project will combine the statistics of the texts and passages quoted in the largest philosophical corpus extant from antiquity (Commentaria in Aristotelem graeca or CAG) with a qualitative and conceptual analysis of selected passages.

Supervisors: Pieter d'Hoine and Orly Lewis

The Presence of Classics in Early Modern Book History

Jonas Fischer, University of Helsinki

The aim of this PhD project is to study the reuse of Latin scientific and technical texts in (Early) Modern England. The Helsinki Computational History Group has used automated methods to detect all the instances of text reuse in the largest collections of printed data for British books (1470–1800, > 250.000 works) available in machine-readable form: Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO-tcp phase II). The result is a massive dataset of linked text-reuse fragments, which are qualified, clustered and compared. Based on this dataset, the project will focus on the text reuse of a selected group of Latin scientific authors such as Lucretius, Pliny the Elder, and Vitruvius, and their reception and reuse in Early Modern works.

Supervisors: Mikko Tolonen and Margherita Fantoli

**Contextual scientometrics: uncovering and understanding referencing patterns to the ancient canon
in modern scholarly discourses**

Luisa Ripoll-Alberola, University of Leipzig

In modern academia, famous authors, thinkers and philosophers from the ancient world continue to shape scholarly discourses. However, precisely which canonical references we find in specific disciplines, and how often they are referred to, is yet to be investigated. The specific objectives of this PhD project are therefore to build a large, multi-disciplinary corpus of scientific journals from JSTOR. Next we will reuse existing computational workflows from the Trismegistos project (TM) to detect ancient canonic references in this corpus. The main analytical objective will be to analyse the extracted references in a scientometric way, allowing us to investigate which of the canonic thinkers are quoted the most in different disciplines and whether there are any diachronic trends to be found. Furthermore, we are interested in co-citation networks, which will help us to understand which canonic authors are frequently mentioned together and may be even forming canonic clusters. In addition to these more traditional scientometric approaches, a core methodological objective is to add a distant reading perspective to the analyses, which allows us to put the canonical references in context and better understand their usage. With this project we develop a novel approach called ‘contextual scientometrics’.

Supervisors: Manuel Burghardt and Mark Depauw

Citations and quotations in the *Naturalis Historia*: creating the canon in the *Encyclopaedia*

Valeria Irene Boano, KU Leuven

The general objective of this PhD project is to study how people are cited in Pliny the Elder’s *Naturalis Historia* (NH), considered the first *Encyclopaedia* of the Ancient World. A monumental collection of 37 books, the *Naturalis Historia* is populated by multiple people who are mentioned in very different roles: sources, witnesses, protagonists of episodes or significant discoveries. Indices of current critical editions represent a useful but relatively flat tool to extrapolate information about the people mentioned: by developing a digital annotation of books 2–6, the PhD candidate will be able to provide an informed picture of the people associated to astronomical and geographical knowledge, and the textual strategies used by Pliny to present them. The specific objectives of this project are (1) to systematically gather quantitative and qualitative information on the people mentioned in NH 2–6; (2) to detect the patterns used by Pliny when selecting the people in relation to the topics treated & structure of the explanations, and thus (3) to understand the mechanisms that guided the representation of sources, witnesses and protagonists in the first attempt to systematize the broad area of “natural sciences” in western history of knowledge.

Supervisors: Margherita Fantoli and Monica Berti