BERT's Conceptual Cartography: Mapping the Landscapes of Meaning

Nina Haket and Ryan Daniels
University of Cambridge
{nch35, rkd43}@cam.ac.uk

Abstract

We present a method for analysing context-sensitive word meanings using BERT embeddings and Gaussian Mixture Models in the fields of lexical pragmatics and Conceptual Engineering. Our methodology generates visual *conceptual landscapes* that reveal how words cluster in different contexts, demonstrated through a case study examining the term PLANET. We provide quantitative metrics for meaning stability and contextual variation, useful for researchers studying lexical pragmatics and meaning change. We also provide an open-source tool which offers an accessible interface for generating visualisations and metrics, requiring minimal technical expertise. Results show that even seemingly straightforward terms exhibit complex meaning landscapes that resist simple definition, highlighting the importance of context-sensitive analyses, combining quantitative metrics and qualitative approaches. This work bridges theoretical pragmatics and computational linguistics, offering empirical grounding for studying how word meanings shift across contexts.

1 Introduction

Language is a complex, dynamic system, constantly evolving and adapting to the contexts in which it is used. Words are not static entities but are deeply embedded in networks of meaning, influenced by both linguistic and extra-linguistic factors. This variability in meaning has long been of interest to linguists, especially in the context of polysemy, the phenomenon of words having multiple related senses (e.g. *paper* as a physical object vs. a scholarly article), and modulations (Recanati, 2010), whereby contextual factors fine-tune a word's interpretation without generating a discrete sense (e.g. an ATM *swallowing* a credit card). We refer to the combination of these polysemous senses and modulation as *contextual meaning variation*, a category encompassing both stable sense multiplicity and more fluid, context-dependent interpretive shifts.

Contextual meaning variations are not merely theoretical concerns – they have significant implications for real-world applications. Conceptual Engineering (CE) is one such domain that directly engages with these issues. CE is concerned with identifying and addressing deficiencies in how words are used, including issues such as vagueness, ambiguity, and biases that distort clear communication (Cappelen and Plunkett, 2020; Cappelen, 2018). Much attention in CE is given to 'improving' words in isolation, but the challenge of modifying word meanings is complicated by the very nature of words: they exist within networks of meanings that shift across different contexts.

In this paper, we propose an interdisciplinary approach that bridges CE, lexical pragmatics, and computational linguistics. We create a tool and method that helps address the practical challenges faced by those navigating the complexities of lexical meaning (e.g. conceptual engineers) by leveraging natural language processing (NLP) techniques to map the intricate relationships within word meanings designed to be broadly useful for researchers in semantics and pragmatics.

Specifically, we use language models such as BERT (Devlin et al., 2019) to generate contextualised embeddings for a selection of words frequently targeted by conceptual engineers, drawn from the spoken component of the British National Corpus 2014 (Love et al., 2017). Using Gaussian Mixture Models (GMMs), we analyze these embeddings to uncover how words cluster in different contextual settings, allowing us to visualise and understand the *conceptual landscapes* of words – how meanings interconnect and shift based on context. These visualisations and metrics map the intricate landscape of meanings associated with a lexical item. Unlike traditional corpus methods such as collocation analyses, our approach con-

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 182-198.
Eugene, Oregon, July 18-20, 2025

denses embeddings into clear visual representations, highlighting the proximity, distinctness, and relationships between meanings while accounting for contextual and distributional complexities. By mapping the conceptual landscapes of words, we offer lexical pragmaticists and conceptual engineers a way to approach the delicate task of understanding contextual variations with greater precision, while simultaneously advancing the capabilities of NLP research to handle complex, context-dependent word meanings. This includes applications in word sense disambiguation (WSD) and dialogue systems.

2 Related Work

While this tool and methodology have wide-ranging applications, we focus on CE as a case study. CE is inherently practical, aiming to actively modify word meanings rather than merely theorising about them. This dimension makes it even more crucial to have robust methods that allow for precise, context-aware revisions to word meanings, ensuring that any interventions are both effective and sensitive to the complexities of language.

2.1 Conceptual engineering

CE is an emerging area of analytic philosophy concerned with improving the tools we use to think and communicate, namely, our words and concepts, when these are found to be defective in some way (Cappelen, 2018; Koch et al., 2023; Isaac et al., 2022). These 'defects' may be theoretical (e.g. vague, misleading, or imprecise terms) or socio-political (e.g. terms that encode harmful ideologies). A prominent example is Haslanger (2000), who argues that biologically grounded definitions of terms like WOMAN and RACE should be replaced with socially grounded ones to better reflect structural realities and serve emancipatory goals. In this sense, CE is a normative project.

Here, we provide empirical tools that can be used by CE practitioners, and, crucially, also by those who wish to critique or scrutinise their efforts. If CE is to be practised at all, it should be done with a full understanding of how meanings actually function across different contexts of use. This paper seeks to separate diagnosis from prescription, and this is where linguistic analysis has a crucial role to play. We offer a method for mapping the actual complexity of word usage, making it possible to ask more informed questions about what kind of change is feasible, who it affects, and where resistance might arise. For a more nuanced discussion of these facets, see Haket (forthcoming). In this sense, the framework is not a blueprint for linguistic intervention, but a diagnostic system for meaning dynamics.

2.2 Lexical Pragmatics

Lexical pragmatics is concerned with how meaning is shaped by context, particularly the influence of pragmatic factors such as speaker intent, discourse context, and social norms. Meaning can vary significantly across different contexts, with words taking on multiple meanings depending on their use. Polysemy has been a key focus in pragmatics, with scholars like Grice (1989) and relevance theorists (Wilson and Sperber, 2006) exploring how contextual cues guide these inferences on the utterance level, and lexical semanticists/pragmaticists exploring the potential for these contextual meaning variations on a lexical level (e.g. Del Pinal 2015).

CE has often treated meanings as fixed, dictionary-style entries that can be revised in isolation (Cappelen, 2018). However, psycholinguistic research shows that understanding speaker meaning in everyday discourse frequently bypasses full semantic decoding (Gibbs, 1984; Gibbs and Moise, 1997; Bezuidenhout and Cutting, 2002). This suggests that CE should shift its focus from static semantic definitions to the dynamic, context-sensitive meanings that arise in real-world use (Pinder, 2020). However, these present a fundamental challenge that has been undertheorised in the CE literature. Utilising this insight means that conceptual engineers must consider not only stable semantic meanings of words but also the ways in which meaning shifts across contexts, through polysemy or through processes like narrowing, broadening, and metaphorical extension. By incorporating contextual meaning variations into CE, we can more precisely map how word meanings function across discourse and avoid overly simplistic or static revisions

2.3 Computational Lexical Pragmatics

If conceptual engineers indeed need to shift their focus to these lexical pragmatic meanings, they need a way of accessing, analysing, and understanding them. After all, these kinds of meanings may not necessarily appear in dictionaries. The challenge lies in systematically analysing how words are actually used across different contexts, a task that has traditionally been difficult to approach at scale. However, recent advances in computational linguistics, particularly through word embeddings like BERT (Devlin et al., 2019), have revolutionised the study of meaning variation. BERT's contextual embeddings have been shown to capture distributional patterns in language, aligning with the American branch of distributionalism (esp. Harris 1954) that semantically similar words tend to occur in similar contexts (Chiang and Yogatama, 2023; Ferret, 2021). BERT's ability to learn such patterns through its masked language modeling objective has revolutionised our ability to study meaning variation.

More specifically, the clustering and analysis of these kinds of embeddings have led to impressive results in a variety of tasks, particularly WSD (Soler and Apidianaki, 2021). BERT embeddings can capture both contextual variations, with the spatial location of embedded words shifting based on their surrounding context (Coenen et al., 2019), and semantic distinctions between different word meanings and usages (Erk and Chronis, 2022; Chronis and Erk, 2020). This dual capability is supported by multiple empirical findings: embeddings of non-polysemous words show higher similarity than polysemous words (Cevoli et al., 2023; Wilson and Marantz, 2022), and BERT's clustering results correlate strongly with human judgments about meaning similarities (Soler and Apidianaki, 2021). BERT can also capture various other linguistic phenomena including metaphorical uses, syntactic roles, and constructions (Giulianelli et al., 2020).

2.4 Aims of this research

Our work makes a threefold contribution to the field. First, we shift the focus of conceptual engineering from static, dictionary-style definitions to the dynamic, context-dependent variations in meaning that arise in discourse, emphasising the importance of lexical pragmatics for conceptual revision. Second, we apply well-established computational lexical tools, such as embedding and clustering techniques, to conceptual engineering, demonstrating how these methods can identify meanings that need revision based on empirical, context-sensitive data. Third, we provide a practical tool for both conceptual engineers and researchers in lexical pragmatics, enabling the analysis of meaning variation in context and helping to identify inconsistencies or ambiguities. By integrating pragmatic theory with computational techniques, our approach allows for a more systematic analysis of both stable meanings and context-dependent shifts, making the revision process more aligned with pragmatic understanding.

3 Methods

In this section, we present a brief overview of the data used, and the computational methods.

3.1 Data

The Spoken British National Corpus (BNC) consists of 1,251 anonymised, unscripted, face-to-face conversations recorded from 672 volunteers from a range of socioeconomic and demographic backgrounds designed to be a representative sample of the British population (Love et al., 2017). The conversations were collected from 2012 to 2014 in a variety of contexts, including business meetings and radio phone-ins, and therefore are representative of everyday vernacular speech. Work on spoken language is underrepresented in previous empirical work on CE, despite it being the primary mode of communication. As such, we chose to focus our research on this area. The Spoken BNC is released under the Spoken BNC2014 User Licence for non-commercial research and teaching purposes.

3.2 Contextual embeddings

BERT (Devlin et al., 2019) is a widely used transformer-based language model, trained on masked token prediction and next-sentence likelihood. Unlike generative models, BERT is bidirectional, attending to both preceding and following tokens. We use the 336M parameter *bert-large-uncased* model, chosen for its balance of performance, efficiency, and simplicity in analysing semantic meaning in the Spoken BNC. BERT's low-resource, low-complexity nature makes it ideal for researchers with limited computational power, to complete our method in under 24 hours. BERT is released under an Apache 2.0 license.

BERT generates *contextual embeddings*, unique embeddings for each token based on its context, in contrast to *static embeddings* like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), which provide a single global representation of a word, ignoring local context. As has been noted, this makes BERT particularly suitable for investigating lexical pragmatic effects: BERT cap-

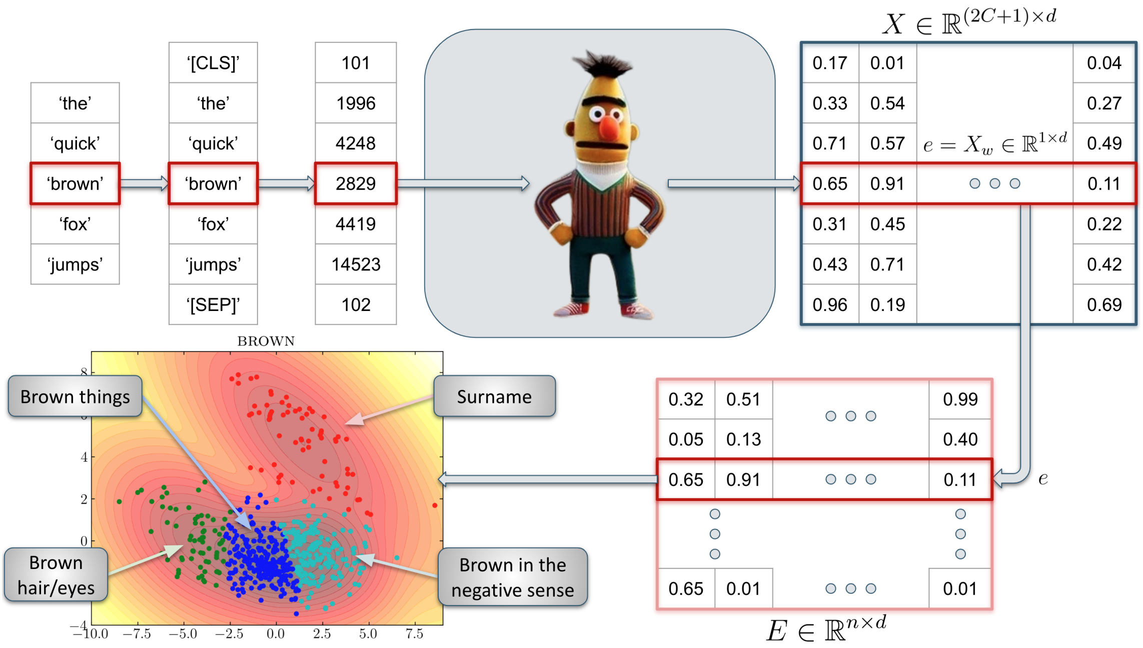

Figure 1: An example of how the target word BROWN is turned into a contextual embedding, e. For a target word the C tokens before and after w are input to BERT. The final embedding e for the target word is then the w^{th} row of the embedding matrix X output from the final hidden layer. A collection of embeddings taken from n sentences are then collated into the matrix E, which is then reduced to 2D and fitted to a GMM.

tures contextual nuances, while static models abstract away this variability.

We generate contextual embeddings for 24 words, *target tokens*, that occur within the Spoken BNC, including words commonly targeted by conceptual engineers such as DUTY, PLANET, TRUTH, and FAMILY (for a full list see Appendix C). These were chosen due to their significance for CE, which usually targets social, moral, political, or philosophical meanings.

We define the context window, C, as half the total number of tokens in the input, excluding the target token, T_w. For a single occurrence of the target token in the text, the total number of tokens fed into BERT is then $2C+1$, where T_w is the middle token: $[T_1, ..., T_C, T_w, T_{C+2}, ..., T_{2C+1}]$. BERT therefore takes as input a $2C+1$ length utterance. The last layer hidden-state is taken as the output – an embedding matrix $X \in \mathbb{R}^{(2C+1)\times d}$. The word contextual embedding is then the w^{th} row, $e = X_w \in \mathbb{R}^{1\times d}$. For n *separate* occurrences of that target token within the text can be represented by the occurrence matrix $E \in \mathbb{R}^{n\times d}$.

3.3 Conceptual landscapes

A Gaussian Mixture Model (GMM) is a method of modelling multimodal data using a combination of K unimodal distributions. We use a GMM to perform unsupervised soft clustering on the embedding matrix E after dimensionality reduction with principal component analysis (PCA). We optimise K and the number of principal components for each word using the Silhouette score (Rousseeuw, 1987). We then perform a robustness analysis using the Adjusted Rand Index (ARI) (Rand, 1971). The ARI measures the similarity between two sets of cluster assignments. Practically, the ARI ranges between [0,1] with 0 indicating entirely random assignments, and 1 indicating perfect agreement between the two cluster assignments. We fix the number of principal components, and then use 1000 random initialisations for training the GMM. The ARI is calculated for all pairs of cluster assignments for the 1000 random initialisations. We calculate the ARI with (i) 2 principal components, and (ii) the optimal number of principal components. The final labels are calculated by aggregating the results of the 1000 runs into a consensus matrix and using hierarchical clustering on this consensus matrix.

To construct the conceptual landscapes we use the GMM fit to the first two principal components with the optimal number of clusters, and find the log-likelihood scores over a defined space (Figure

1). Limitations and ethical considerations of this methodology can be found in Appendices A and B.

3.4 Metrics

We use four main metrics to describe the landscapes: *maximum explained variance* (MEV), *self-similarity*, *intra-group similarity*, and *inter-group similarity*. The definitions used here closely follow those from Ethayarajh (2019).

MEV If target token T_w appears in sentence i then e_i is the corresponding embedding. The values $\sigma_1, ..., \sigma_m$ are then the first m singular values of the centered occurrence matrix. The MEV is the proportion of variance explained by the first principal component, given by

$$\text{MEV}(w) = \frac{\sigma_1^2}{\sum_i \sigma_i^2} \tag{1}$$

and ranges over $[0, 1]$. MEV indicates the extent to which a contextual embedding could be replaced by a static embedding. If MEV is high it means that the first principal component alone accounts for most of the variation in how a word is used across all its different contexts in the corpus. Even though a given model (such as BERT) produces different embeddings for a word in each context, these embeddings are not scattered randomly. Instead, their differences lie mostly along a single primary axis of meaning variation. You could therefore, in principle, project all the contextual embeddings onto this single line with relatively low loss of information about their overall distribution. A word with a high MEV therefore indicates a uniform consistency of word usage (for example, if the word BARK is always used in the context of “like a dog”). Conversely, if the MEV is low, then no one vector can adequately capture to variation in usage. In terms of CE then, the MEV measures the extent to which changing the semantic meaning is likely to influence speaker meanings.

Self-similarity The self-similarity is the average cosine similarity between embedding vectors, given by

$$\text{Sim}(w) = \frac{1}{n^2 - n} \sum_i \sum_{j \neq i} \cos(e_i, e_j) \tag{2}$$

and ranges over $[0, 1]$. For CE, this metric gives a value of *how much* variation we see within the word. A word with a high self-similarity is constrained in its diversity of usage and meaning, whereas a low self-similarity indicates high diversity in usage.

Anisotropy (the non-uniform distributions of words in embedding space) in LLM contextual embeddings is well documented (Ethayarajh, 2019). It is therefore necessary to control for anisotropy by taking a random sample of embeddings and finding the total average similarity. This baseline is then subtracted from the similarities for each word.

Intra-group similarity Let $e_{k,i}$ be the embedding e_i assigned to label k with n_k members. The global average intra-group similarity for K groups is then

$$\text{Intra} = \frac{\sum_k \sum_i \sum_{j \neq i} \cos(e_{k,i}, e_{k,j})}{\sum_k (n_k^2 - n_k)} \tag{3}$$

For CE, this metric measures similarity within assigned contextual clusters. If the clusters contain contextually similar usages, this score should be high. A high intra-group similarity suggests that the word is used consistently within each cluster, facilitating more precise and effective CE interventions. This allows for targeted modifications to the word’s meaning and usage, making it easier to implement changes and achieve the desired conceptual clarity.

Inter-group similarity Let $e_{k,i}$ be the embedding e_i assigned to label k, where n_l are those embeddings *not* assigned to label k. The global average inter-group similarity for K groups is then

$$\text{Inter} = \frac{\sum_k \sum_{l \neq k} \sum_i \sum_j \cos(e_{k,i}, e_{l,j})}{\sum_k \sum_{l \neq k} n_k n_l} \tag{4}$$

For CE, this metric compares members of a single contextual cluster with members from *other* contextual clusters. If the clusters are contextually different from one another, and each individual cluster contains usages which are contextually similar, this score should be low. High inter-group variation suggests more distinct boundaries between contexts, delineating specific usages, which can make CE easier to implement since it can target specific contexts without interference from others.

3.5 Tool

To facilitate practical application of this methodology, we have made a tool publicly available at`https://github.com/acceleratescience/`

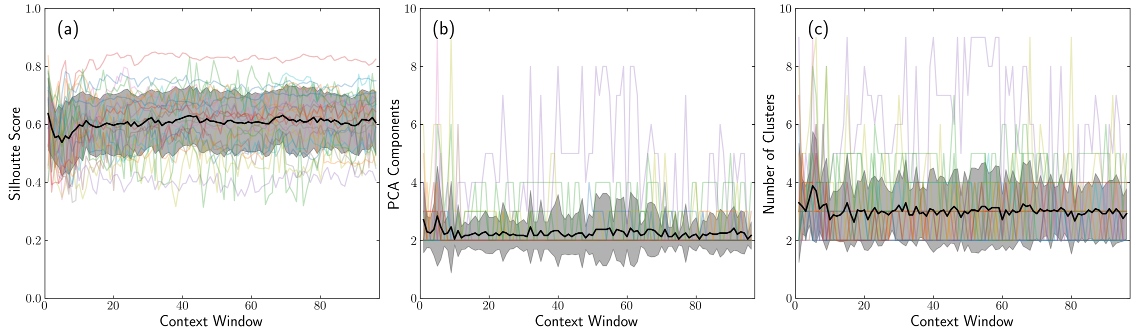

Figure 2: The Silhouette scores (a), optimal number of principal components (b), and optimal number of clusters (c) for each GMM fit to each word. Bold lines indicate averages, and shaded regions indicate the standard deviation.

conceptual-cartography. The tool provides an intuitive interface for generating conceptual landscapes and computing the metrics described in this paper. Conceptual engineers can input their target words and corresponding text corpora to visualise meaning clusters, analyze contextual variations, and quantify polysemy through our suite of metrics (MEV, self-similarity, intra-group and inter-group similarity). This enables precise identification of meaning variations and supports evidence-based decision-making in conceptual revision projects. The tool includes comprehensive documentation and example analyses, making it accessible to researchers regardless of their computational background.

4 Results and Discussion

We applied our methodology to a range of words commonly targeted by conceptual engineers, spanning scientific terms (e.g., WEIGHT, ENERGY, PLANET), philosophical concepts (e.g., TRUTH, FREEDOM, KNOWLEDGE), social constructs (e.g., FAMILY, MARRIAGE, EDUCATION), and terms related to technology (e.g., COMPUTER). A complete list of words analysed can be found in Appendix C, and presentation of all the calculated metrics for each word can be found in Table 1 and Table 2.

4.1 Context size

Figure 2 shows the result of optimising the GMM for (a) Silhouette scores, (b) number of principal components, and (c) number of clusters for different context window sizes for the target words. Note that the minimum value of the Silhouette score is achieved at $C = 4$, and therefore when the total number of tokens is ~ 9. The utterance lengths of the Spoken BNC are approximately power-law distributed (see 9) with an average utterance length of ~ 10. This suggests that taking a single utterance as input to BERT may be insufficient to capture the full contextual meaning of the target word. This lends credence to modern approaches to meaning that emphasise meaning across entire discourses as opposed to within a single utterance (Jaszczolt, 2015). As the total number of input tokens exceeds the average utterance length, the Silhouette score increases quickly and remains relatively steady, achieving a maximum at $C \sim 40$.

Importantly, the average number of optimal principal components across words and context windows is ~ 2, and the optimal number of principal components is 2 for every word, except for DUTY, and MARRIAGE. For the following sections, we choose a context window of 40, where the Silhouette score is at a maximum. For all subsequent analyses, the number of clusters is fixed to the optimal number of clusters for each word (for Silhouette scores, optimal principal components and optimal number of clusters for each word, see Figure 8).

4.2 Cluster properties

Figure 3 shows the MEV scores and average self-similarities after correcting for anisotropy (a), and the intra-group similarity and inter-group similarity (b) for the target words. These results are in strong agreement with Ethayarajh that static embeddings would be poor substitutes for the contextual embeddings obtained from BERT. In addition, we also found that a control for anisotropy was not necessary when reducing dimensions.

Figure 3c shows that there is an excellent agreement between the ARI scores when using 2 principal components and when using the optimal number of components, suggesting that the 2D representations capture a substantial amount of the clus-

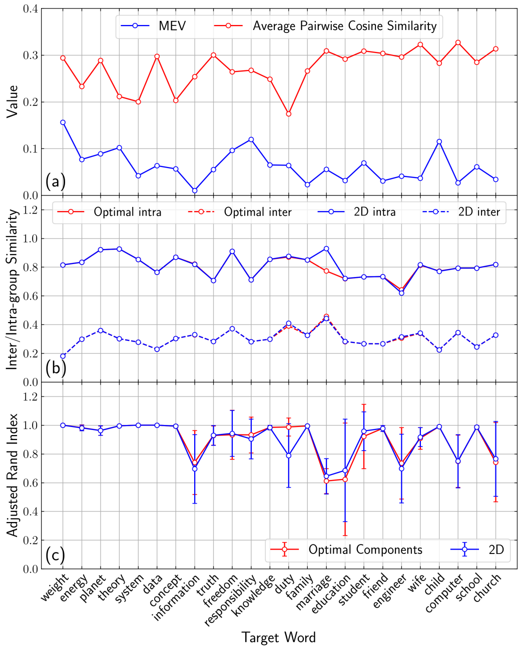

Figure 3: (a) Anisotropy-corrected self-similarity (red) and maximum explained variance (blue). (b) Intra-(solid line) and inter-group (dashed line) similarity for the optimal number of principal components (red), and for 2 principal components (blue). (c) ARI for 1000 GMMs fitted to the optimal number of principal components (red), and for 2 principal components (blue). Error bars are the standard deviations.

tering structure found in the higher-dimensional space. Secondly, the ARI scores show significant variability across words. Words such as WEIGHT, SYSTEM, and FAMILY have high average ARI, and low variance; words such as INFORMATION, EDUCATION, and DUTY, have lower average ARI and higher variance.

Words with high ARIs cluster consistently across different initialisations, indicating a well-defined, stable model, and therefore a well-defined and stable conceptual landscape. The contexts are likely to be more distinct and less ambiguous. Words with lower ARIs may have more ambiguous or varied contexts, causing the clusters to overlap. Therefore, the varying levels of stability reflect the differences between contextual distinctions and ambiguity. The ARI scores for each word are understandably correlated with the Silhouette scores ($r = 0.723$, $p < 0.0001$), given both metrics aim to quantify a measure of cluster quality and stability albeit from different perspectives.

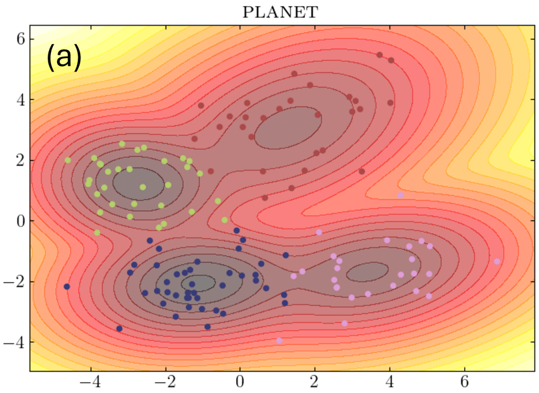

Figure 4: The conceptual landscapes generated using the negative log-likelihood of the GMM predictions in 2D for PLANET with 4 clusters.

4.3 Conceptual landscapes

Since the average number of optimal principal components is approximately 2, it is therefore reasonable to use the 2D conceptual landscape as an indicator of contextual word usage without significant information loss. Figure 4 shows example conceptual landscapes for PLANET (for all target words and landscapes, Figures 6 and 7).

4.3.1 Planet

Due to space constraints and the fact that conceptual engineers typically focus on refining meanings of individual words or closely related sets, this paper analyses a single term (PLANET) to demonstrate how empirical methods can inform CE. The redefinition of PLANET by the IAU in 2006, particularly the exclusion of Pluto, is one of the most frequently mentioned case studies in CE (Landes and Reuter, 2024). Here, it serves here not as a diachronic case study of semantic change, but as a touchstone for the challenges conceptual engineers face when revising the meanings of contextually variable terms. We examine the current semantic landscape in which such revisions take place. Specifically, we ask: when a formal body like the IAU proposes a revision, what kind of semantic structure is it intervening in—and what does that structure imply about the likely uptake, resistance, or diffusion of the revised meaning?

Our analysis reveals both stability and complexity in how PLANET is used. The high ARI of 0.96 indicates consistent, clearly identifiable usage patterns, suggesting distinct meanings that conceptual engineers could potentially target. However, the low MEV of 0.09 demonstrates that no single, static representation can capture the term's full range of uses. The self-similarity score of 0.29, while relatively high, points to considerable contextual variation. Together, these metrics suggest that PLANET exists in a complex semantic space with multiple distinct but related meanings.

This complexity is further illuminated by our identification of four distinct clusters of usage through Gaussian Mixture Model (GMM) analysis and qualitative interpretation:

1. **Astronomical:** Used in scientific contexts to describe celestial bodies in space.
2. **Environmental:** Used in discussions about global ecology or climate change, such as 'saving the planet'.
3. **Metaphorical:** Used to describe a person or object as alien or incomprehensible, as in 'from another planet'.
4. **Hyperbolic:** Used in casual or media contexts to exaggerate the scope of issues or concepts, as in 'worst thing on the planet'.

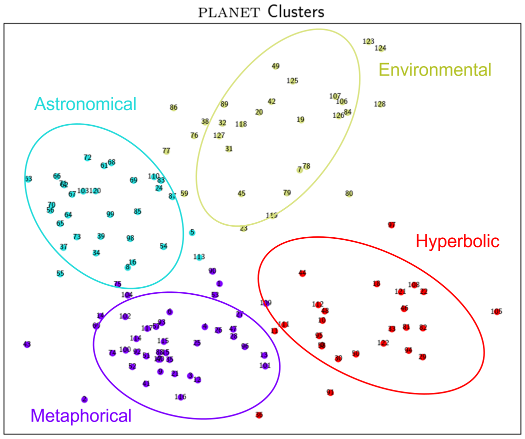

Figure 5: Clusters for PLANET after qualitative analysis.

The PLANETexample illustrates several critical insights for CE.The IAU's redefinition assumes a clear boundary between the astronomical meaning of PLANET and its other uses, such as in environmental or metaphorical contexts. As such, the use of PLANET in environmental contexts ('saving the planet', or even the phrase 'the planet') is of no consequence, as this definition does not depend on whether dwarf planets are PLANETS or not. However, our empirical findings suggest that these meanings are not as easily separated as this theoretical model suggests.

These clusters are not isolated silos: intra-cluster similarity is high (0.92), but inter-cluster similarity remains non-trivial (0.36) indicating gradience and potential overlap between uses. This matters for CE, because it undermines the assumption that a revision to one sense (e.g. the astronomical sense targeted by the IAU) can be neatly isolated from others (e.g. the environmental or metaphorical ones). For instance, even if 'the planet' in 'save the planet' refers to Earth rather than any celestial body, our analysis shows that it remains semantically entangled with the broader category of PLANET.The variability across these different clusters of meaning (especially the overlap between the environmental and metaphorical senses) illustrates the importance of understanding modulation for CE. If conceptual engineers attempt to modify a word's meaning in one context, the resulting revision can inadvertently affect other uses, complicating the task of meaning modification.The observed gradience in meaning—where senses overlap and shift between contexts—illustrates a core challenge for CE. If one sense is revised without accounting for these overlapping uses, unintended consequences may arise in contexts that seem unrelated at first glance, undermining the intended revision.

This complexity is what conceptual engineers must reckon with. Rather than assuming that a term like PLANET can be revised in one domain (e.g. astronomy) without consequence, our data suggests that contextual variations make such revisions porous. In short, if CE is to intervene effectively, it must first understand the semantic terrain it is operating within—and our metrics offer a scalable, replicable way to map that terrain.

4.4 Usage in Conceptual Engineering and Beyond

Conceptual landscapes offer significant theoretical and practical advantages for conceptual engineers. By visualising the variations in meaning of a term like PLANET, conceptual engineers can pinpoint the kinds of meaning they aim to revise and assess how it interacts with other meanings, helping to identify overlaps, dependencies, and links. For instance, revising the astronomical sense of PLANET might clarify scientific discourse, but without careful consideration, it could unintentionally disrupt the metaphorical or environmental uses prevalent in public discussions. These landscapes provide a framework for addressing meaning with precision, sensitivity, and empirical grounding, without requiring extensive training in computational techniques, embeddings, or computer science.

Our methodology offers concrete benefits for CE practice specifically through a structured approach across all stages of the process (see e.g. Koch et al. 2023):

Diagnostic Phase: Identify major meaning clusters, quantify stability (MEV/self-similarity), and map relationships between senses/modulations (inter-cluster similarity).

Planning Phase: Target clusters for revision, predict interference with others, and identify optimal intervention points in the meaning network.

Implementation Phase: Monitor meaning shifts, assess uptake in target contexts, and identify unintended consequences in related clusters.

This framework shifts CE from intuition-based practice to an empirically-grounded methodology, enabling practitioners to visualise and quantify conceptual landscapes. Our tool makes this approach accessible to conceptual engineers without computational expertise, bridging the gap between theoretical CE and practical application. By providing a data-driven understanding of polysemy and variation, it supports both CE and lexical pragmatics. The methodology combines CE's focus on individual words with NLP's large-scale analysis, allowing researchers to explore both the nuances of specific words and broader linguistic landscapes with greater precision.

5 Conclusion

This study introduces a novel methodology for analysing context-sensitive word meanings, bridging the fields of CE, lexical pragmatics, and computational linguistics. First, we have argued for shifting the focus of CE from static definitions to dynamic, context-sensitive meanings. Second, we have provided a methodology for conceptual engineers and lexical pragmaticists to apply computational tools to map the conceptual landscapes of words, revealing polysemy and contextual variations.

As demonstrated through our analysis of PLANET, our approach can effectively identify distinct meaning clusters while quantifying their relationships. The four identified senses (astronomical, environmental, metaphorical, and hyperbolic) and their associated metrics (ARI of 0.96, MEV of 0.09, indicating consistent clustering and strong context-dependence) demonstrate how words can have clearly identifiable yet interrelated meanings that resist simple definition. By leveraging BERT embeddings and Gaussian Mixture Models (GMMs), we generate conceptual landscapes that visualise meaning variation and provide quantitative metrics such as MEV and self-similarity.

Finally, we have created an accessible toolkit that provides a practical and systematic framework for conceptual engineers, linguistic theorists, and others to analyse meaning variation and guide meaning revision efforts, empowering researchers to base their analyses on empirical data rather than abstract intuition.

Acknowledgments

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). RKD acknowledges funding from the Accelerate Programme for Scientific Discovery Research Fellowship. NH acknowledges funding from the Cambridge Trust's International Scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the funders. The authors thank Aditya Ravuri, Catherine Breslin, and Kasia Jaszczolt for useful discussions and feedback.

References

Annelie Ädel. 2010. Using corpora to teach academic writing: Challenges for the direct approach. In *Corpus based approaches to ELT*, pages 39–55. Bloomsbury Publishing.

Jack Bandy and Nicholas Vincent. 2021. Addressing" documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Anne Bezuidenhout and J Cooper Cutting. 2002. Literal meaning, minimal propositions, and pragmatic processing. *J. Pragmat.*, 34(4):433–456.

Herman Cappelen. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press, Oxford.

Herman Cappelen and David Plunkett. 2020. *Introduction: A Guided Tour of Conceptual Engineering and Conceptual Ethics*, pages 1–34. Oxford University Press.

Benedetta Cevoli, Chris Watkins, Yang Gao, and Kathleen Rastle. 2023. Shades of meaning: Uncovering the geometry of ambiguous word representations through contextualised language models. *arXiv preprint arXiv:2304.13597*.

Ting-Rui Chiang and Dani Yogatama. 2023. The distributional hypothesis does not fully explain the benefits of masked language model pretraining. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it‘s like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715.

G. Del Pinal. 2015. Dual content semantics, privative adjectives, and dynamic compositionality. *Semantics and Pragmatics*, 8(7):1–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katrin Erk and Gabriella Chronis. 2022. Word Embeddings are Word Story Embeddings (and That's Fine). In *Algebraic Structures in Natural Language*. CRC Press.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics.

Olivier Ferret. 2021. Using Distributional Principles for the Semantic Study of Contextual Language Models. In *https://aclanthology.org/events/paclic-2021/*, https://aclanthology.org/events/paclic-2021/, pages 189–200, Shanghai, China.

Raymond W. Gibbs. 1984. Literal meaning and psychological theory. *Cognitive Science*, 8(3):275–304.

Raymond W. Gibbs and Jessica F. Moise. 1997. Pragmatics in understanding what is said. *Cognition*, 62(1):51–74.

M. Giulianelli, M. Del Tredici, and R. Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973. Association for Computational Linguistics.

H. P. Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Nina Haket. forthcoming. Navigating meaning spaces: A contextualist approach to conceptual engineering. Manuscript in progress.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2–3):146–162.

Sally Haslanger. 2000. Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs*, 34(1):31–55.

Manuel Gustavo Isaac, Steffen Koch, and Ryan Nefdt. 2022. Conceptual engineering: A road map to practice. *Philosophy Compass*, 17(10):1–15.

Kasia M. Jaszczolt. 2015. Default Semantics. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 743–770. Oxford University Press, Oxford.

Steffen Koch, Guido Löhr, and Mark Pinder. 2023. Recent work in the theory of conceptual engineering. *Analysis*, page anad032.

Ethan Landes and Kevin Reuter. 2024. Conceptual revision in action. Preprint.

Robbie Love, Claire. Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha. Association for Computational Linguistics.

Mark Pinder. 2020. Conceptual engineering, speaker-meaning and philosophy. *Inquiry*, pages 1–15.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

François Recanati. 2010. *Truth-Conditional Pragmatics*. Oxford University Press, Oxford, New York.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

A. Garí Soler and M. Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Deirdre Wilson and Dan Sperber. 2006. Relevance theory. *The handbook of pragmatics*, pages 606–632.

Kyra Wilson and Alec Marantz. 2022. Contextual embeddings can distinguish homonymy from polysemy in a human-like way. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 144–155, Trento, Italy. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Limitations

BERT is pretrained on BOOKCORPUS (Zhu et al., 2015) and English WIKIPEDIA (Devlin et al., 2019), which may introduce biases reflective of these contexts into our analysis. By adjusting for anisotropy, we mitigate some of these biases. However, this is not a complete solution. Future work should explore other models and fine-tune on more diverse datasets. In addition, the Spoken BNC includes speech from British individuals over a limited time period, which may not reflect contemporary language use and perspectives, and does not encompass linguistic data from other countries.

While 2D projections are useful for visualising and comparing word contexts, there are instances where higher-dimensional embeddings (e.g., for MARRIAGE) provide a clearer representation of semantic differences. This highlights a limitation of our current approach, as projecting down to 2D may obscure important nuances. Future work should explore higher-dimensional embeddings and non-linear dimensionality reduction techniques (e.g., t-SNE, UMAP) to aid visualisation.

Corpus linguistics has been critiqued for its 'inevitable focus on surface forms' (Ädel, 2010), risking an impoverished view of language. We acknowledge this limitation, but argue that CE, being applied and practice-oriented, benefits from observational data on how words are used in context.

B Ethical Considerations

B.1 Use of Language Models

Cultural and language bias. BERT's training data contains cultural biases, including problematic content and skewed religious representation (Bandy and Vincent, 2021). These may affect downstream tasks. Our framework may help identify such biases in training corpora.

The predominance of English in training data limits cultural representation. Fine-tuning on more diverse datasets could mitigate inequities in downstream applications.

Environmental impact. We opted to use BERT for its relative efficiency and smaller environmental footprint, in contrast to larger language models.

Privacy and copyright. While BERT's sources (English Wikipedia, BOOKCORPUS) reduce some privacy concerns, the latter was scraped without author consent, raising ethical issues about data usage.

B.2 Conceptual Engineering

CE attempts to reshape meanings, which can appear overly prescriptive. As meanings are bound to culture and identity, changes not inclusive of diverse perspectives risk alienating the communities they aim to help.

Moreover, CE projects can have social or political ripple effects. We therefore emphasise that this paper offers a descriptive tool: it does not advocate for any particular conceptual change. We provide data about current usage, without prescribing what words should mean.

C Full List of Tested Words

The tested words are:

- *weight*
- *energy*
- *planet*
- *theory*
- *system*
- *data*
- *concept*
- *information*
- *truth*
- *freedom*
- *responsibility*
- *knowledge*
- *duty*
- *family*
- *marriage*
- *education*
- *student*
- *friend*
- *engineer*
- *wife*
- *child*
- *computer*
- *school*

Conceptual landscapes for all words are provided in Figure 6 and Figure 7.

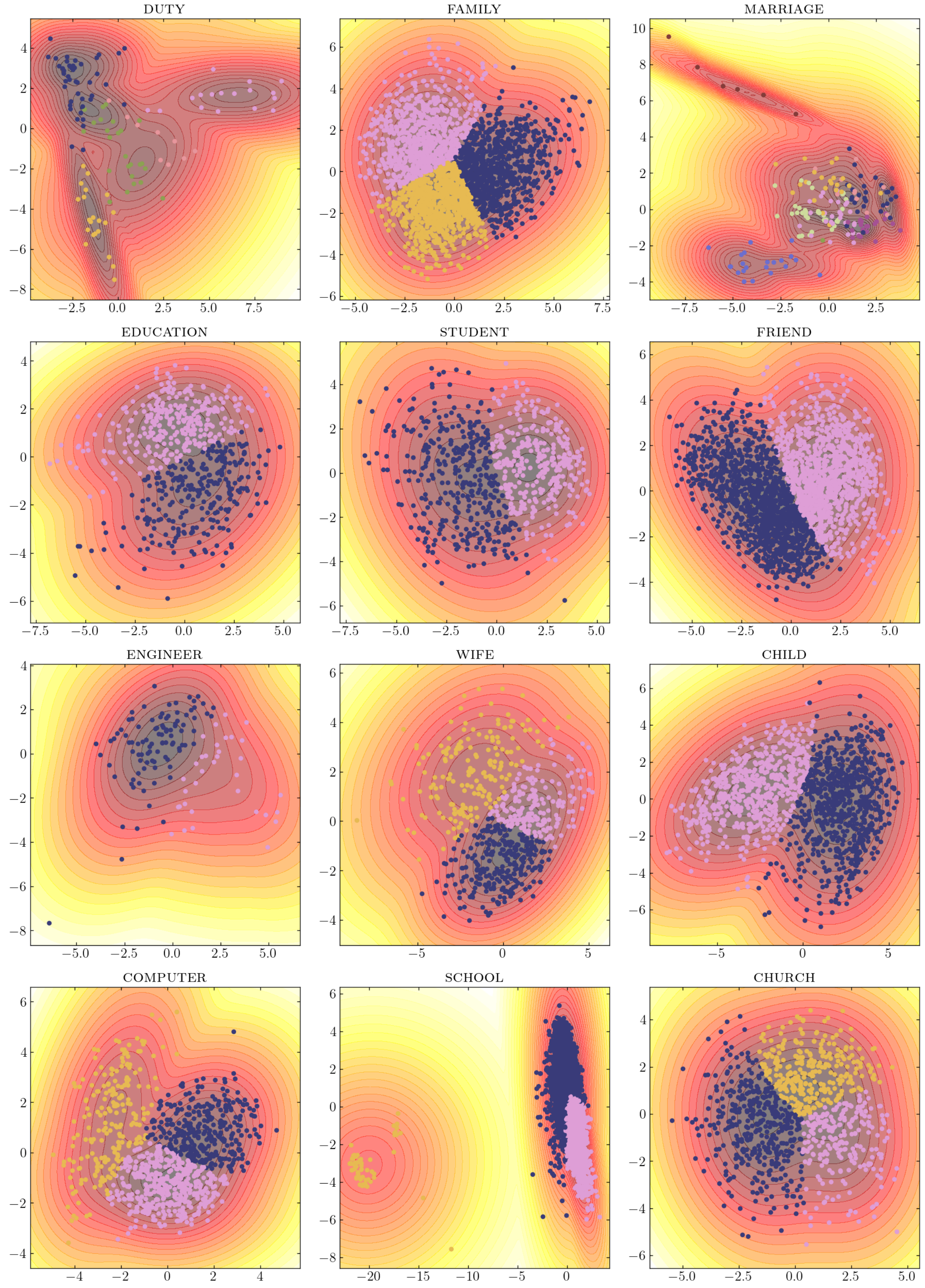

Figure 6: Consensus cluster maps (negative log-likelihood of GMM predictions) for DUTY through CHURCH.

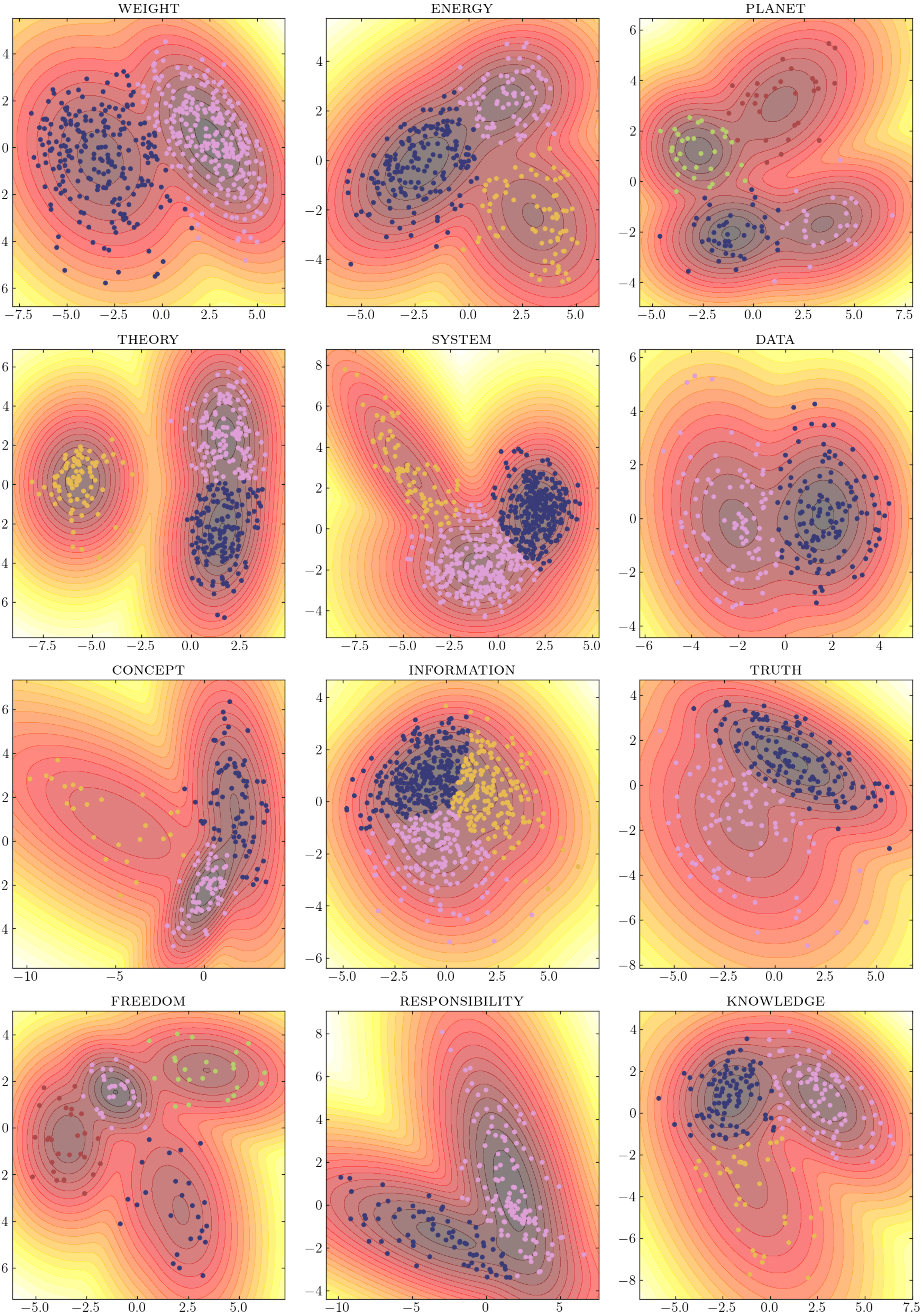

Figure 7: Consensus cluster maps (negative log-likelihood of GMM predictions) for WEIGHT through KNOWLEDGE.

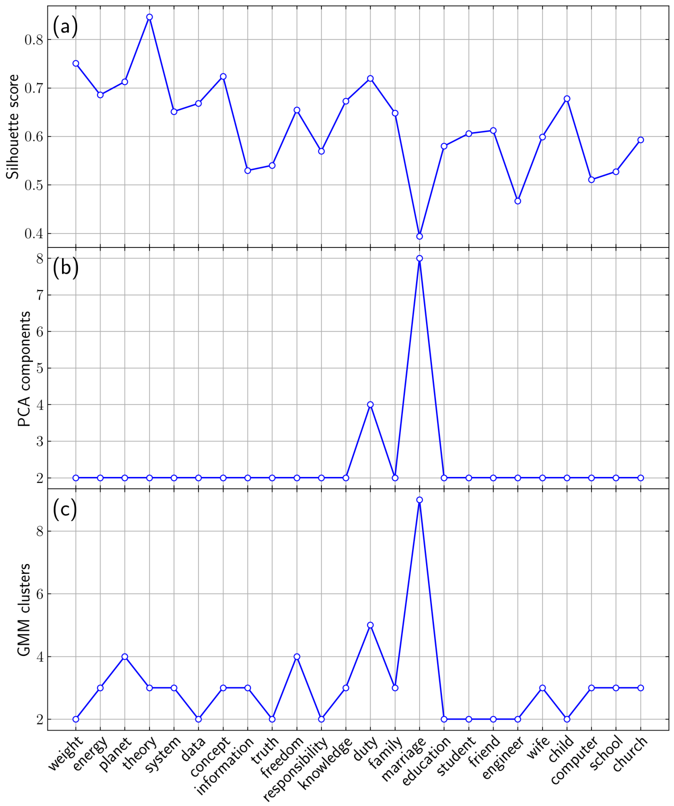

Figure 8: Hyperparameter optimization results: (a) Silhouette scores, (b) number of principal components, (c) number of clusters. Silhouette and ARI scores are closely correlated.

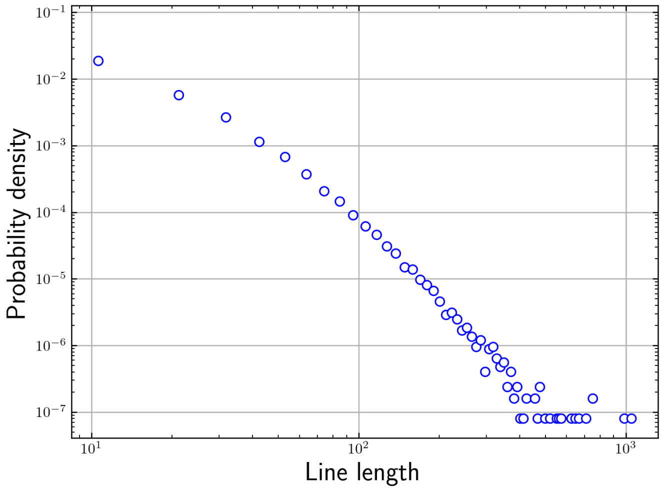

Figure 9: Distribution of utterance lengths in the Spoken BNC. These follow a power-law distribution, with an average of 10 words per utterance.

Word	Optimal Components	Optimal Clusters	Best Score	Self- Similarity	MEV	Optimal ARI	Optimal ARI std	2D ARI	2D ARI std
weight	2	2	0.75065166	0.29663867	0.15589406440022405	1.0	0.0	1.0	0.0
energy	2	3	0.68554884	0.2355035	0.0764136765566644	0.9825799112461127	0.01962176431331328	0.9810035825407845	0.019943568812842878
planet	2	4	0.7128142	0.2913559	0.0882366070740109	0.9623680945153172	0.033088163527237174	0.9626463129788648	0.033075721085404144
theory	2	3	0.84671956	0.21386349	0.10176437079286657	0.9954917857014244	0.004464708197051879	0.9955062177600033	0.004464824823707685
system	2	3	0.6512991	0.20273864	0.04158690442568404	1.0	0.0	1.0	0.0
data	2	2	0.66792816	0.3001163	0.0629035355143655	1.0	0.0	1.0	0.0
concept	2	3	0.723671	0.20564383	0.056091484246831205	0.9927769304661705	0.010173316215442398	0.9935736885883468	0.009859014275290734
information	2	3	0.5296078	0.25662804	0.009946750868021742	0.7406537395085958	0.2229459439592183	0.6960789526143788	0.23915597814367864
truth	2	2	0.54026866	0.30280912	0.054876043198309576	0.9280861918891818	0.06604710754854408	0.9294042503312504	0.06944080370936948
freedom	2	4	0.65434194	0.26655453	0.09563030806831296	0.9334675211205564	0.1704048934448	0.9427797663532299	0.15956748076657506
responsibility	2	2	0.56938255	0.2700225	0.11916091303327889	0.9322590837578956	0.12430886548367147	0.904437876644496	0.13800775307016663
knowledge	2	3	0.6726736	0.25102633	0.0645349155540209	0.9843930515547132	0.015854844628385403	0.9841431866165496	0.015856813383450186
duty	4	5	0.7196939	0.17664373	0.06377767425054486	0.9874551707150581	0.06252875440515283	0.7889657682855794	0.22171666221553094
family	2	3	0.64838034	0.26847154	0.022635997134030736	0.9943836321000173	0.005135851540094211	0.994669251883583	0.004894600248935083
marriage	8	9	0.3940875	0.31149036	0.05500554472960928	0.6112942192156345	0.08721537746728976	0.6447443787878189	0.12265073796622647
education	2	2	0.57998776	0.29428303	0.031460283242946446	0.6239074938781386	0.39106286425586695	0.6854307222784576	0.35662533525437484
student	2	2	0.60584253	0.31139386	0.06900681973577344	0.9219265130861394	0.22414619352034612	0.9590315441198952	0.1349587985075622
friend	2	2	0.6121608	0.30614358	0.03020277056434878	0.9774785414648027	0.021040128884952422	0.9774135569385687	0.02137824727231833
engineer	2	2	0.466553	0.29832488	0.04074953892079064	0.7353913559980679	0.24881847024839246	0.6982006637956051	0.23926154624159524
wife	2	3	0.5990231	0.32535738	0.03624206212147031	0.9089545688639884	0.07546172732778701	0.9178098402172326	0.06638190772251286
child	2	2	0.6781394	0.28536147	0.11498878751245134	0.9901787999403814	0.007877072801540326	0.9904813585767728	0.007709889184639891
computer	2	3	0.5107223	0.32953215	0.026559695009549786	0.7491521491427756	0.18228835360128462	0.7498979023596061	0.18482899770396627
school	2	3	0.52753365	0.28709567	0.060625261536956604	0.9864745788983414	0.011797563212586125	0.9867719126782513	0.011881742870975943
church	2	3	0.5926918	0.31584865	0.03375614676062694	0.7429283349034067	0.2752246758009878	0.7661161198120406	0.25985354840800257

Table 1: Calculated metrics for 24 target words using dimensionality reduction and unsupervised clustering. Metrics include the number of optimal principal components and clusters, best clustering score, self-similarity, maximum explained variance (MEV), ARI scores and standard deviations for both optimal clustering and 2D projections.

Word	Optimal Intra-Sim	Optimal Inter-Sim	2D Intra-Sim	2D Inter-Sim
weight	0.8156033219962284	0.18104519595828528	0.8156033219962284	0.18104515090349865
energy	0.8336862218346286	0.29917733958914533	0.8336863203976584	0.29917730863040515
planet	0.9207608160844708	0.3597553812266942	0.9207608160844708	0.35975542080850764
theory	0.9263468231635071	0.30127710391438284	0.9263466688882307	0.30127714540843425
system	0.8536792740152507	0.27814700771867	0.8536796265171682	0.2781468876547384
data	0.7635351625646621	0.22914614096660874	0.763537767362795	0.2291443617544815
concept	0.8683227585248771	0.30286055940233236	0.8683227585248771	0.30286050245991253
information	0.821972462161749	0.3293009304867715	0.819028850508441	0.3297536590393733
truth	0.7062944748230764	0.28244020454910296	0.7062943393117325	0.28244022355133097
freedom	0.9097507468259896	0.3721182697521081	0.9097507468259896	0.3721182697521081
responsibility	0.7112177734375	0.2816186389568326	0.7112175071022727	0.28161882269883615
knowledge	0.8539526334736376	0.29840904028655746	0.8539525793884304	0.2984091032783977
duty	0.8704321464283045	0.39059547301983527	0.8753667447726858	0.40915019581755635
family	0.8506438458340466	0.32548975138527925	0.850408401614284	0.3256464671847802
marriage	0.7729137680385885	0.4574278943574383	0.9296340574523867	0.443919260225337
education	0.7194930980302446	0.28191425273944803	0.7206263273206777	0.28236683933054896
student	0.7317915722548086	0.2667373108328637	0.7317915722548086	0.26673728025891486
friend	0.7342716988092002	0.2663353340758285	0.7342721319883346	0.2663351627458536
engineer	0.6412440521413054	0.3058740765440698	0.6201696425980734	0.3141900634765625
wife	0.8135365350376823	0.33955498015490126	0.8165262413059602	0.3409786710666057
child	0.7725739291386711	0.2235273103563482	0.7711400170618056	0.22415934626025402
computer	0.7929313357494175	0.3450574308027275	0.7929309680417951	0.34505763451584726
school	0.7931535947179521	0.24504607627722041	0.793153645676212	0.24504624575719003
church	0.818619789088437	0.3265899456336431	0.818619789088437	0.3265899456336431

Table 2: Calculated metrics for 24 target words using dimensionality reduction and unsupervised clustering. Metrics include the Inter-Similarity and Intra-Similarity for both optimal clustering and 2D projections.