Decomposing Co-occurrence Matrices into Interpretable Components as Formal Concepts

Anonymous ACL submission

Abstract

This study addresses the interpretability of word representations through an investigation of a co-occurrence matrix. Employing the mathematical methodology of Formal Concept Analysis, we reveal an underlying structure that is amenable to human interpretation. Furthermore, we unveil the emergence of hierarchical and geometrical structures within word vectors as consequences of word usage. Our results reveal a significant correspondence between semantic and vector spaces.

1 Introduction

Word vector representations are central to natural language processing, as they capture semantic and syntactic features (Lenci, 2018). They are used as input for Transformer-based language models (Vaswani et al., 2017; Devlin et al., 2019), where static embeddings are contextualized. Although word representations are effective when performing tasks, the interpretability of their dimensions remains an active research topic (Şenel et al., 2018). Levy and Goldberg (2014a) found neural word embeddings to be uninterpretable while acknowledging that sparse vectors capture some latent topics. Geva et al. (2022) pioneered efforts to interpret dynamic embeddings in GPT-2 (Radford et al., 2019) by projection into the vocabulary space. However, the systematic interpretation remains an open issue.

Many studies have revealed a certain correspondence between a vector space of word representations and its respective semantic space of word meanings. The most well-known example of such correspondence is the parallelogram formed in the vector space by the embeddings of words in analogical relations (e.g. *king:queen::man:woman*) (Mikolov et al., 2013c). Other semantic relationships also exhibit geometrical counterparts, such as semantic composition with vector addition (Mikolov et al., 2013b; Mitchell and Lapata, 2008), hypernymy captured by linear projection (Fu et al., 2014), and polysemy as a linear combination of vectors (Arora et al., 2018). Regarding the theoretical analysis of embeddings, Levy and Goldberg (2014b) suggested that word2vec (Mikolov et al., 2013a) is equivalent to the factorization of a word co-occurrence matrix. Arora et al. (2016) proposed a generative model in which PMI-based word embeddings exhibit linear structures. These studies collectively hint that the latent structure in the matrix reflects linguistic regularities and is inherently embedded within vector representations. Therefore, understanding the word co-occurrence matrix represents a cornerstone in elucidating the interpretability of word representations.

Unlike prior studies, we directly address the mathematical structure of a word co-occurrence matrix. To achieve this, we focused on the correspondence between a vector space and a semantic space. Specifically, we used Formal Concept Analysis (FCA), a field of applied mathematics (Ganter and Wille, 2012), to formally characterize the internal structure of a matrix. We claim that *a formal concept* mathematically defined in the word co-occurrence matrix corresponds to interpretable categories. Furthermore, we demonstrate that a hierarchical structure of formal concepts emerges as a geometric formation in the vector space.

Our contributions are threefold. First, we propose two methods that apply FCA to real-valued data: binarization by varying thresholds, and fuzzification of FCA. Second, we empirically show that the formal concepts in the word co-occurrence matrix coincide with interpretable categories. Third, we present a novel algorithm to detect formal concepts, which is capable of disambiguating polysemous words.

To our knowledge, this is the first study to apply FCA to a word-word co-occurrence matrix. Our study offers a new approach to uncover linguistic structures, bridging the gap between semantic cognition and mathematical representation.

2 Formal concept analysis of word co-occurrence matrix

2.1 Basics of FCA

FCA is related to order theory and abstract algebra. It mathematizes *concepts* and *conceptual hierarchy* (Ganter and Wille, 2012). A concept comprises a pair of its extents (objects) and its intents (attributes). Concepts can form a hierarchy known as a *lattice*. FCA has been empirically applied for data mining and ontology (Poelmans et al., 2013), especially in bioinfomatics (Roscoe et al., 2022).

A **formal context** $\mathbb{K} := (G, M, I)$ consists of two sets G, M and a binary relation $I \subseteq G \times M$. The elements of G and M are called **objects** and **attributes**, respectively. For $g \in G$ and $m \in M$, a relation $(g, m) \in I$ means that the object g has the attribute m. We define two derivation operators; $\uparrow$: $2^G \rightarrow 2^M$ maps a subset of objects to a subset of attributes, and its reverse $\downarrow$: $2^M \rightarrow 2^G$ maps attributes to objects. For $A \subseteq G, B \subseteq M$,

$$A \uparrow := \{m \in M \mid (g, m) \in I \ (\forall g \in A)\} \quad (1)$$

$$B \downarrow := \{g \in G \mid (g, m) \in I \ (\forall m \in B)\} \quad (2)$$

$A \uparrow \subseteq M$ is the set of attributes common to all objects in A, whereas $B \downarrow \subseteq G$ is the set of objects that possess all the attributes in B. It can be shown that $A \subseteq B \downarrow \Leftrightarrow B \subseteq A \uparrow$, which is a structure-preserving (order-reversing) correspondence between ordered sets known as a Galois connection (Davey and Priestley, 2002).

A **formal concept** of the context (G, M, I) is defined as a pair $(A, B) \in 2^G \times 2^M$ where both $A \uparrow = B$ and $B \downarrow = A$ hold. A and B are considered the extent and intent, respectively, of the concept (A, B). The compositions of two derivation operators $\uparrow\downarrow$: $2^G \rightarrow 2^G$ and $\downarrow\uparrow$: $2^M \rightarrow 2^M$ are closure operators (Davey and Priestley, 2002), with a formal concept defined as the fixed point of these operations. If a formal context is represented as a binary matrix, it corresponds to a maximal rectangular (submatrix) with all ones in its entries when the rows and columns are appropriately reordered.

A formal concept can also be equated with a maximal **biclique**, i.e., a complete subgraph of a bipartite graph (Chiaselotti et al., 2015). All elements of A and B are completely connected within that subgraph.

2.2 Rational and benefit of using FCA

A word co-occurrence matrix, used as input data to learn word embeddings, is constructed by counting the frequency of a target-context word pair that co-occurs in the neighborhood. By regarding target words as objects and context words as attributes, we can express this co-occurrence as a binary relation. Thus, we can treat a co-occurrence matrix as a formal context.

FCA is effective in analyzing co-occurrence matrices for three reasons. First, it can characterize a local structure within the matrix. Second, formal concepts can capture relations between more than three words, which cannot be represented by individual pairwise relationships, yielding a richer analysis of the structure. Third, we can define (partial) order relation between formal concepts. A semantic relationship such as hypernymy can be formalized by such an order relation. We further demonstrate the function of FCA in Section 3.

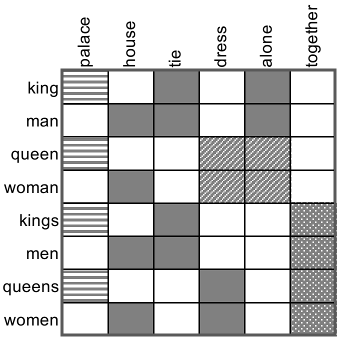
To apply the crisp (binary) FCA to a real-valued co-occurrence matrix, we tested two approaches. First, we simply binarized the matrix values by thresholds, with a varying threshold method deployed to flexibly locate formal concepts (Section 4). Second, we extended the crisp FCA to an FCA built on fuzzy logic (Section 5).

3 Demonstration using synthetic data

3.1 Artificial toy corpus

We examined how FCA handles a word co-occurrence matrix using a toy corpus. We demonstrated that formal concepts capture semantic categories emerging from word usage in the corpus, and introduced a **concept lattice** of FCA to illustrate the hierarchical structure of concepts.

The demonstration contains 1) a corpus of 24 synthetic sentences with 17 words (Appendix A), 2) a co-occurrence matrix obtained from the corpus, and 3) word vectors acquired from the matrix (Fig. 1). The corpus is designed to replicate a geometric formation of the analogy relation. Specifically, we targeted eight words—*king, queen, man, woman*, and their plurals—so that their vectors formed a parallelepiped. The sentences were expressed analogously: E.g., *'king (queen) live in palace"*, whereas *"man (woman) live in house"*. The co-occurrence matrix $X \in \{0, 1\}^{17\times17}$ is binary, where $X_{ij} = 1$ if two words co-occur in a sentence and $X_{ij} = 0$ otherwise. Each row of this matrix represents a word vector. Projected on the



	palace	house	tie	dress	alone	together
king	▤		■		■	
man		■	■		■	
queen	▤			▨	▨	
woman		■		▨	▨	
kings	▤		■			░
men		■	■			░
queens	▤			■		░
women		■		■		░

Figure 1: Binary co-occurrence (sub)matrix: Each entry is 1 if shaded and 0 otherwise. Each row is a word vector. Three submatrices with shade patterns indicate formal concepts f_i, e_j, v_k.

3-dimensional space, the eight word vectors form a parallelepiped (Fig. 2).

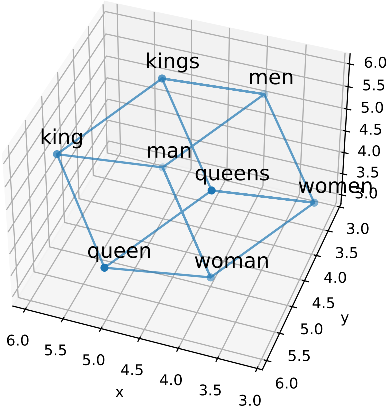


Figure 2: A parallelepiped emerges when eight word vectors (rows) are projected onto 3-dimensional space.

3.2 Detecting formal concepts

We now apply FCA to the matrix X. Although formal concepts can be determined by applying the closure operator $\uparrow\downarrow$, a simplified method is to find a rectangular in the matrix. For example, the submatrix of rows $i \in \{1, 3, 4, 7\}$ and columns $j \in \{1\}$ represents a formal concept, as all of its entries are 1s and no other rectangular matrix contains it. This concept represents a pair of the extent $\{king, queen, kings, queens\}$ and the intent $\{palace\}$, interpreted as "royal."

There are a total of 28 formal concepts in this matrix (see Appendix B for the list and notation). They are classified into five types, including two trivial ones wherein one element is empty. Examples of the three non-trivial types include the following:

$$f_1 := (\{king, man, kings, men\}, \{tie\}) \quad (3)$$

$$e_1 := (\{king, man\}, \{tie, alone\}) \quad (4)$$

$$v_1 := (\{king\}, \{tie, palace, alone\}) \quad (5)$$

To see hierarchical relations between formal concepts, we first define the order relation. Let $\mathfrak{B}(G, M, I)$ be the set of all concepts of (G, M, I). Given $(A_1, B_1), (A_2, B_2) \in \mathfrak{B}(G, M, I)$,

$$(A_1, B_1) \leq (A_2, B_2) \overset{\text{def}}{\Longleftrightarrow} A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2 \quad (6)$$

Thus, if the extent A_1 is contained by the extent A_2, then the formal concept (A_1, B_1) is less than or equal to (A_2, B_2). Owing to the Galois connection, $A_1 \subseteq A_2$ holds if and only if $B_1 \supseteq B_2$. Then, $\langle \mathfrak{B}(G, M, I) : \leq \rangle$ is a complete lattice known as a **concept lattice**, a nonempty ordered set where a join and a meet exist for all elements and subsets. Fig. 3 visualizes all ordered relations between the formal concepts identified in the matrix X. We observe that the lattice of formal concepts (Fig. 3) corresponds to the parallelepiped (Fig. 2). This suggests that geometric relations between word vectors reflect the hierarchical structure latent in the word co-occurrence matrix.

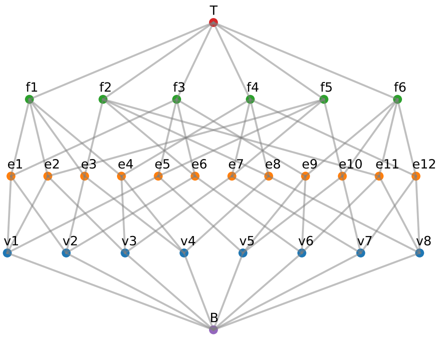


Figure 3: Lattice of formal concepts. Each node represents a formal concept. The nodes correspond to geometric simplices of the parallelepiped: 8 vertices, 12 edges, 6 faces.

3.3 Three implications of FCA

First, FCA allows us to easily interpret the identified formal concepts. For example, f_1 should be labeled as **masculine** from its extent {*king, kings, man, men*}, whereas f_6, with the extent {*queen, queens, woman, women*}, must be labeled as **feminine**. The other f-type concepts can be labeled as **royal**, **common**, **singlular** and **plural**. Thus, formal concepts coincide with semantic concepts.

Second, the formal concept v_1 (*king*) can be seen as the intersection of three others—f_1, f_3, f_5—analogous to a vertex included in three faces. Semantically, *king* is something royal, masculine, and single. This relation can be algebraically formulated as $v_1 = f_1 \wedge f_3 \wedge f_5$ where $\wedge$ is a *meet* operation.

Third, pairs of opposing faces in the parallelepiped form complementary concepts such as **masculine** vs. **feminine**. Mathematically, we can construct a *formal concept algebra* by defining additional operations as axioms (Wille, 2004). Using this algebra, the formal concept of **masculine** can be demonstrated to complement that of **feminine**; $\neg f_1 = f_6$ where $\neg$ is a negation. Thus, we can characterize antonymy as an algebraic complement.

In summary, the co-occurrence matrix exhibits the geometrical and algebraic structures formed by interpretable formal concepts, revealing a structural correspondence between the semantic and vector spaces.

4 Experiment 1: FCA by binarization

4.1 Applying FCA to a real data

We now demonstrate that formal concepts can be defined on actual word co-occurrence data and correspond to both semantic and syntactic categories. We compared two methods that apply the crisp FCA to a real-valued matrix: binarization by thresholding, and Fuzzy FCA.

For binarization, we apply a certain threshold to a value in each entry of the co-occurrence matrix. We adopted PPMI (positive point-wise mutual information; Niwa and Nitta, 1994) as it yields the best results in the semantic task (Bullinaria and Levy, 2012). We found it to be more useful to flexibly adjust a threshold that is locally determined in the PPMI matrix, rather than applying a single value.

4.2 Algorithm to identify formal concepts

We designed a novel algorithm to locate formal concepts through the conversion of two derivation operators (Eq. 1 and 2). The corresponding pseudo-algorithm is shown in Algorithm 1. Given a PPMI co-occurrence matrix X and set of target words S as a seed, the algorithm returns a formal concept $(S\uparrow\downarrow, S\uparrow)$, which is a pair of two subsets of the vocabulary. Here, $S\uparrow\downarrow$ is the closed set of S.

Algorithm 1 Varying Threshold Method

```
Input: $X \in \mathbb{R}^{N\times N}, S := \{w_i\}_{i\in I_S}, k \in \mathbb{N}$
Output: $FC := (S\uparrow\downarrow, S\uparrow), t \in \mathbb{R}$
 1: function FINDFORMALCONCEPT($S, k$)
 2:     for $j \leftarrow 1$ to $N$ do
 3:         $m_j \leftarrow \min_{i\in I_S} X_{ij}$
 4:     end for
 5:     Sort $[m_j]$ in descending order $\leftarrow [m_{p(j)}]$
 6:     $J_{S\uparrow} \leftarrow \{p(j)\}_{j\le k}$
 7:     $S\uparrow \leftarrow \{w_j\}_{j\in J_{S\uparrow}}$
 8:     $t \leftarrow m_{p(k)}$
 9:     $I_{S\uparrow\downarrow} \leftarrow \emptyset$
10:     for $i \leftarrow 1$ to $N$ do
11:         $\mu_i \leftarrow \min_{j\in J_{S\uparrow}} X_{ij}$
12:         if $\mu_i \ge t$ then
13:             $I_{S\uparrow\downarrow} \leftarrow I_{S\uparrow\downarrow} \cup \{i\}$
14:         end if
15:     end for
16:     $S\uparrow\downarrow \leftarrow \{w_i\}_{i\in I_{S\uparrow\downarrow}}$
17:     return $(S\uparrow\downarrow, S\uparrow), t$
18: end function
```

The first derivation operator $\uparrow$ must identify context words that co-occur with all target words in S. In other words, a context word is selected when it has all PPMI values exceeding the threshold t for the target words in S. Equivalently, any PPMI value that the seed words have with the context word should not be less than t, meaning that their minimum must be greater than or equal to t. As indicated in Line 3, the algorithm finds the minimum value that the seed words (in rows $\forall i \in I_S$) have against a certain context word (in a column $j \in \{1, \ldots, N\}$), sorts them in descending order (Line 5), and selects the first k context words (columns) $S\uparrow$ (Line 6). The threshold is automatically determined as the kth largest minimum value (Line 8). Next, an inverse operation executes. Given $S\uparrow$, the algorithm finds a minimum value over the context words $S\uparrow$ ($J_{S\uparrow}$ in the column index) against a target word in a row i (Line 11) and selects the target words (rows $I_{S\uparrow\downarrow}$) with minimum values exceeding the threshold (Line 13), which form $S\uparrow\downarrow$.

$I_{S\uparrow\downarrow}$ and $J_{S\uparrow}$ are subsets of rows and columns corresponding to $S\uparrow\downarrow$ and $S\uparrow$, respectively. t is the determined threshold. The algorithm ensures that a submatrix $(X_{ij})_{i\in I_{S\uparrow\downarrow}, j\in J_{S\uparrow}}$ satisfies:

$$X_{ij} \ge t \quad (i \in I_{S\uparrow\downarrow},\, j \in J_{S\uparrow}) \tag{7}$$

$$X_{ij} < t \quad (\forall j \notin J_{S\uparrow}, \exists i \in I_{S\uparrow\downarrow}) \tag{8}$$

$$X_{ij} < t \quad (\forall i \notin I_{S\uparrow\downarrow}, \exists j \in J_{S\uparrow}) \tag{9}$$

Note that the submatrix of $I_{S\uparrow\downarrow} \times J_{S\uparrow}$ is discriminated from its neighbouring area. Its inner region has higher values than t (Eq. 7), whereas each of its exterior rows and columns horizontally (Eq. 8) and vertically (Eq. 9) adjacent to the submatrix contains at least one cell below the threshold.

4.3 Semantic categorization test

The following experiment was conducted to verify that the formal concepts identified from the co-occurrence matrix coincide with interpretable categories.

Test set We adopted two existing test sets from Lindh-Knuutila and Honkela (2015) containing semantic categories: the Battig set (Bullinaria and Levy, 2012), comprising 53 categories with 10 words for each, and BLESS (Baroni and Lenci, 2011), containing 17 categories with 5-17 words for each. We also compiled two additional sets: Series and Syntactic. The categories tested are listed in Appendix D.

Procedure For each category, we systematically furnished the algorithm with all possible word pairs as seeds derived from the category's word set. Next, we identified the optimal seed that yields the most extensive set of accurately classified words. We then assessed how effectively the algorithm retrieves the correct words from the optimal seed for the given category (**Precision, Recall**). Because the word sets are not necessarily exhaustive, we also regarded those missed words as correct, based on our human judgement (**Extended precision**)[1].

Data The co-occurrence matrix, constructed from the English Wikipedia dump (20171001)[2], spanned 2.9B tokens, counted with a window size of 10 and converted into PPMI values. To keep the matrix size manageable, we limited the vocabulary to the 10K most frequent words.

4.4 Results

Qualitative results Table 1 presents output samples produced by the algorithm. When given {*large, huge*} as a seed, the algorithm returned {*large, huge, enormous, vast*} as the extent and {*sums, amounts, quantities*} as the intent, which constitutes a formal concept. All PPMI values within this concept exceeded 3.95. This formal concept can be labeled as "largeness" or adjective of size, which implies that it is indeed interpretable. Similar results held for other seeds.

Quantitative results Table 2 shows that 61.5–84.3% of the identified extent words matched the category labels in the test sets (**Extended precision**). Furthermore, 56.3–76.8% of the words in the test sets were retrieved by the algorithm (**Recall**). Semantic categories in Battig, BLESS, and Series were more effectively captured by formal concepts than syntactic categories. We also observed that homogeneous categories (e.g., Country) frequently formed formal concepts.

4.5 Analysis

The results suggest that formal concepts overlap with interpretable categories. Furthermore, we find it intriguing that FCA mathematically characterizes the internal structures of the matrix. Recall that higher PPMI values discriminate the submatrix of a formal concept from its neighbors, forming a local plateau-like structure that is not necessarily captured by the cosine similarity. This insight offers two use cases for the proposed algorithm.

Disambiguating polysemy A target word can participate in multiple formal concepts. By inputting seed words with different associations, we found that polysemous words such as *tie* and *spring* have multiple formal concepts, as shown in Table 3. We observed that separate formal concepts (e.g., clothing, match, fasten) may contain the same word (e.g., *tie*) in their extents. Three separate plateaus may share the same row as visualized in Fig. 4.

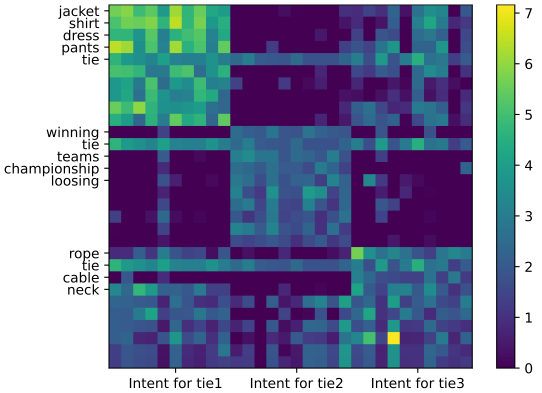


Figure 4: PPMI submatrix of three formal concepts containing the same polysemous word *tie*. For ease of visibility, the row for *tie* is presented multiple times.

Arora et al. (2018) discovered that the embeddings of polysemous words can be decomposed as

[1] The annotation was done by one of the authors, who is non-native but has educational experience in the U.S.

[2] CC BY-SA 3.0; https://dumps.wikimedia.org/legal.html

Seed	**Formal Concept** (upper:extents; lower:intents)	**Th.**	**Category**
large, huge	*large, huge, enormous, vast*	3.95	Adjective of size
	sums, amounts, quantities		
church, temple, mosque	*chapel, church, mosque, synagogue, temple*	2.85	Religious Buildings
	worship, jpg,ruined		
quicker, bigger, warmer	*bigger, brighter, colder, cooler, heavier, hotter, louder,...*	2.45	Comparatives
	than, considerably, deeper		

Table 1: Examples of formal concepts identified from a binarized PPMI matrix. Given seed words, the algorithm returns an extent-intent pair representing a formal concept. The parameter k was set to 3. Th. means threshold.

Test set	**Prec.**	**Ex.prec.**	**Recall**	LKH
Battig	51.0	81.7	64.4	(37.0)
BLESS	57.8	84.3	67.0	(64.7)
Series	62.8	82.7	76.8	-
Syntactic	57.1	61.5	56.3	-

Table 2: Average precision, extended precision, and recall over the categories ($k = 3$), expressed as percentages. LKH lists % of the categories identified by Lindh-Knuutila and Honkela (2015).

linear combinations of sense vectors. Our finding suggest that these vectors reflect separate formal concepts, and that the embeddings inherit the inner structure of the co-occurrence matrix.

Measuring a similarity in subspace The proposed algorithm generates a byproduct that can be used to investigate the relationship between the multiple vectors (the rows of the matrix) in a subspace. By reusing Lines 3–5 in the algorithm, we can determine whether target words in a seed share certain context words in limited dimensions and are semantically related in the shared context.

Specifically, we propose the **subspace similarity** $\phi(S)$ defined as

$$\phi(S) := \frac{1}{k} \sum_{i=1}^{k} m_{p(i)} \tag{10}$$

for a group of words $S = \{w_i\}_{i \in I_S}$, where $m_j := \min_{i \in I_S} X_{ij}$, $p(i)$ is a permuted index in descending order and k is a hyperparameter for the scope of subspace. The notation is the same as in Algorithm 1. The subspace similarity is the mean of the thresholds t determined over different parameter values up to k selected dimensions. Fig. 5 shows the computed values of the subspace similarity for several word groups. These results indicate that semantically related groups share certain context words locally, even if their cosine similarities are low. Generally, randomly chosen vectors in high-dimensional space tend to be orthogonal, which implies a low chance of detecting correlations in any dimension. In contrast, a higher subspace similarity should suggest that a certain structure can be defined more than incidentally.

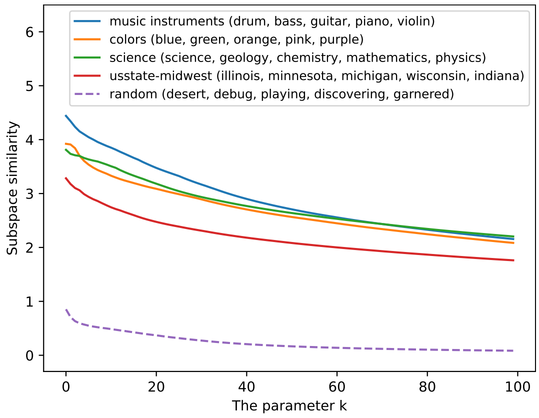


Figure 5: Subspace similarity for groups of five words. Semantically related groups exhibit significantly higher values than the randomly selected set.

5 Experiment 2: Applying Fuzzy FCA

5.1 Fuzzification of FCA

Our second application of FCA to a real-valued matrix involves the fuzzification of the crisp FCA by incorporating fuzzy set theory (Ojeda-Hernández et al., 2023). A fuzzy set formalizes an ambiguous set, such as "a set of tall people," by assigning a degree of membership to each element. In Appendix C, we give the definition of a fuzzy formal concept and show that it is equivalent to a rank-one submatrix under our proposed specification.

The problem of finding fuzzy formal concepts can be regarded as that of identifying nonnegative rank-one submatrices in a PPMI matrix. Because it is NP-hard to exactly decompose a matrix into nonnegative factors (Vavasis, 2010), we obtained an approximation by deploying nonnegative matrix

Word (sense)	**Seed**	**Extent of Formal concept**
`tie1` (clothing)	*tie, pants, shirt*	*collar, jacket, pants, shirt, tie, wears*
`tie2` (match)	*tie, teams, winning*	*championship, playoffs, teams, tie, winning*
`tie3` (fasten)	*tie, cable, rope*	*cable, loose, neck, rope, tie*
`spring1` (season)	*spring, autumn, month*	*autumn, cold, coldest, cooler, dry, month, rainfall,...*
`spring2` (metal)	*spring, wheel, suspension*	*fitted, mounted, rear, spring, suspension, wheel, wheels*
`spring3` (water)	*spring, creek, river*	*basin, brook, creek, reservoir, river, spring, stream*

Table 3: The extent of multiple formal concepts comprises polysemous words. The proposed algorithm is able to disambiguate these contexts in response to the seeds associated with them. The parameter k was set to 5 except for the case of `spring1` ($k = 10$).

factorization (NMF; Lee and Seung, 1999), as its L_1 regularization on factor matrices is considered effective in making them sparse. We controlled the sparseness so that the decomposed submatrices became disjoint.

NMF decomposes $X \in R_+^{m\times n}$ into two matrices $W \in R_+^{m\times r}$ and $H \in R_+^{n\times r}$ so that $X = WH^T = \sum_{k=1}^{r} \boldsymbol{w}_k \boldsymbol{h}_k^T$, where $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ are the kth columns of W and H, respectively. The outer product $\boldsymbol{w}_k \boldsymbol{h}_k^T$ is of rank one and preferably sparse, thereby approximating a fuzzy formal concept. The loss function is

$$\begin{aligned} \mathcal{L}_\alpha(W,H) = \frac{1}{2}\|X - WH^T\|_F^2 \\ +\alpha n\|W\|_1 + \alpha m\|H\|_1 \end{aligned} \tag{11}$$

We recursively applied NMF [3] over three rounds—first to the PPMI matrix as in Section 4, then twice to the positive residual matrices resulting from decomposition—factorizing into $r = 300$ components each round. Parameters for the $L1$ norms were set to $\alpha = 5, 3, 1 \times 10^{-4}$ for each round.

5.2 Results

We manually labeled 900 rank-one submatrices by reviewing the words corresponding to the largest entries in $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ (see Appendix E.1 for details). We then classified the submatrices among four categories to assess how well the labels describe the words in each formal concept[4] (Table 4). More than 95% of the acquired formal concepts were interpretable to some extent.

5.3 Analysis

We found that Fuzzy FCA reveals the same formal concepts as the crisp FCA. For example, the three categories listed in Table 2 also appear as rank-one submatrices. In fact, NMF detected more eligible words (e.g. *immense, massive* for Largeness, *shrine* for Religious Buildings). This observation demonstrates the robustness of FCA, as well as the correlation between the two methods.

Class	**R1**	**R2**	**R3**	LKH
Descriptive	182	75	73	27
Partial	56	63	48	72
Meaningful	56	150	158	2
Nonsense	6	12	21	11
Total	300	300	300	112

Table 4: Decomposed rank-one submatrices in four classes for each round, indicating how the submatrices coincide with labeled categories. Definitions are provided in Appendix E.2 and the numbers under LKH are cited from Lindh-Knuutila and Honkela (2015).

Another interesting finding is that two types of rank-one submatrices were discovered: a *clique* type with identical rows and columns, and a *biclique* type with different rows and columns. An example of the latter is ({*explain, describe, discuss, ...*}, {*beliefs, concepts, ideas,...*}), which represents a verb phrase for an act of communication.

6 Discussion

6.1 Why do formal concepts correspond to intepretable categories ?

As noted in Section 2, a formal concept is equivalent to a biclique, which means that the words in it are densely connected. A group of words form a dense community if the words are repeatedly used together. Furthermore, if the same latent state always emits the same set of words, then those words are repeatedly counted as co-occurrence, thereby forming a formal concept. The random walk model of Arora et al. (2016) captures the same mechanism to generate linearly structured embeddings.

[3] NMF from Scikit-learn library: BSD license.

[4] The same as the footnote 1.

However, a latent state is not necessarily limited to a topic, i.e., a state based on thematic proximity. As revealed in Section 4, formal concepts may reflect functional proximity, e.g. the comparative. Furthermore, we observed phrasal proximity, as in a verb phrase. We can conjecture that words in compound nouns correspond to formal concepts by sequential proximity. Thus, a broad range of semantic and syntactic patterns of word usage may be captured as a formal concept.

These results open questions regarding why the human brain understands and operates formal concepts, which may be the subject of future studies of semantic cognition. Our approach may provide a quantitative method to address these questions.

6.2 How can formal concepts be fully captured ?

We designed two methods that apply FCA to a real-valued matrix to detect interpretable formal concepts, although we do not yet have a theory to assess and relate the methods.

In general, the challenge of FCA in applied studies is scalability stemming from computational complexity, which must be addressed when increasing the size of a co-occurrence matrix. Another challenge is posed by heterogeneous data from large corpora. Specifically, we observed that interpretable formal concepts are detected at different threshold levels (Section 4) and by layered factorization (Section 5). The latent structures at different scales indicate that multiple formal contexts coexist in the matrix as if they were superposed, and they were probably generated separately. Thus, the rank-one submatrices may be disjoint, superposed, or overlapping. To ensure appropriate extraction, the algorithm must depend upon the modeling of generative processes, which is also a topic for a future study.

6.3 What do embeddings represent after all?

Recall that formal concepts defined as rank-one submatrices appear as components of matrix factorization $X = WH^T$ (Section 5). While a column of W corresponds to a fuzzy set that constitutes each formal concept, a row of W is used as a word embedding. Thus, a value in each dimension of the embedding can be seen as the "coordinate" of the corresponding formal concept. The other matrix H is considered to encode attributes. The embeddings, acquired by matrix factorization or implicit factorization (Levy and Goldberg, 2014b), must inherit the structures of formal concepts, as the factor matrices can be mutually transformed.

7 Related studies

Several studies demonstrated that sparse embeddings are interpretable. Murphy et al. (2012) and Biggs et al. (2008) applied nonnegative matrix factorization with a sparsity constraint to word-document co-occurrence data and discovered topics. Other studies (Faruqui et al., 2015; Park et al., 2017; Jang and Myaeng, 2017) investigated word embeddings to restore interpretability by using sparsity. We mathematically formalized the latent structure in the word co-occurrence matrix, which prior studies might have empirically detected.

FCA has been applied in linguistics (Priss, 2005), primarily for ontology. Cimiano et al. (2005) applied FCA for the automatic acquisition of taxonomies from a corpus. Moraes and Lima (2012) built a semantic structure by setting the S-V-C tuples of the annotated corpora as a formal context. Berend et al. (2018) used FCA by binarizing sparse word embedding for hypernymy discovery. In contrast to these studies, we deployed FCA to explore the structure of the matrix itself, which revealed the mathematical correspondence between semantic and vector spaces.

Gastaldi (2021) delved into the underlying mechanism of word embeddings from a linguistic-philosophical perspective and pointed out simultaneous codetermination or *bi-duality* between terms and contexts as a significant feature of language, which we believe to have successfully formalized via FCA. Our mathematical approach to interpreting co-occurrence data may shed light on the structure of language, as Bradley et al. (in press) frames language in category theory .

8 Summary

This study establishes a mathematical characterization of the relationship between semantic and vector spaces, employing FCA to investigate a word co-occurrence matrix. Our experiments demonstrate that identified formal concepts align with interpretable categories. Using synthetic data, we also illustrated the emergence of hierarchical structures from word usage. Subsequent challenges include theoretical sophistication in applying FCA, exploring generative modeling, and delving into cognitive inquiries.

9 Limitations and risks

Our study is inherently exploratory, with the aim of communicating critical insights in a timely manner before exhaustively diving into a comprehensive analysis. Consequently, a more thorough investigation and nuanced analysis are deferred to future work, acknowledging that the current study serves as a preliminary exploration that lays the foundation for deeper scrutiny.

Another limitation of this work stems from the reliance on a singular dataset for our analysis. Although our findings reveal compelling patterns within the chosen dataset, generalizability across diverse data sets remains an unexplored avenue. We anticipate similar trends in other data sets, but a comprehensive cross-validation across various sources is pending. Future research efforts should extend our methodology to encompass a wider spectrum of data sets, ensuring the robustness and applicability of our observed trends across different contexts.

The study constitutes a fundamental analysis aimed at identifying mathematical properties within linguistic statistical data, thus enhancing interpretability. Notably, no significant material risks were identified throughout the investigation and will not be seen due to the nature of the analytical approach employed.

Acknowledgements

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.

William F Battig and William E Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1.

Radim Belohlavek and Vilem Vychodil. 2012. Formal concept analysis and linguistic hedges. *International Journal of General Systems*, 41(5):503–532.

Gábor Berend, Márton Makrai, and Péter Földiák. 2018. 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 928–934, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Biggs, Ali Ghodsi, and Stephen Vavasis. 2008. Nonnegative matrix factorization via rank-one downdate. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 64–71, New York, NY, USA. Association for Computing Machinery.

Tai-Danae Bradley, Juan Luis Gastaldi, and John Terilla. in press. The structure of meaning in language: parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society, Februaly 2024*.

John A Bullinaria and Joseph P Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, 44:890–907.

Giampiero Chiaselotti, Davide Ciucci, and Tommaso Gentile. 2015. Simple undirected graphs as formal contexts. *Formal Concept Analysis: 13th International Conference, ICFCA 2015, Nerja, Spain, June 23-26, 2015, Proceedings 13*, pages 287–302.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of artificial intelligence research*, 24:305–339.

Brian A Davey and Hilary A Priestley. 2002. *Introduction to lattices and order*. Cambridge university press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the*

52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1199–1209, Baltimore, Maryland. Association for Computational Linguistics.

Bernhard Ganter and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

Juan Luis Gastaldi. 2021. Why can computers understand natural language? the structuralist image of language behind word embeddings. *Philosophy & Technology*, 34(1):149–214.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. 2016. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286.

Kyoung-Rok Jang and Sung-Hyon Myaeng. 2017. Elucidating conceptual properties from word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 91–95, Valencia, Spain. Association for Computational Linguistics.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27.

Tiina Lindh-Knuutila and Timo Honkela. 2015. Exploratory analysis of semantic categories: comparing data-driven and human similarity judgments. *Computational Cognitive Science*, 1:1–25.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Sílvia Moraes and Vera Lima. 2012. Combining formal concept analysis and semantic information for building ontological structures from texts : an exploratory study. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3653–3660, Istanbul, Turkey. European Language Resources Association (ELRA).

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012: Technical Papers*, pages 1933–1950, Mumbai.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Manuel Ojeda-Hernández, Domingo López-Rodríguez, and Pablo Cordero. 2023. Fuzzy algebras of concepts. *Axioms*, 12(4).

Sungjoon Park, JinYeong Bak, and Alice Oh. 2017. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411, Copenhagen, Denmark. Association for Computational Linguistics.

Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. 2013. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 40(16):6538–6560.

Uta Priss. 2005. Linguistic applications of formal concept analysis. In *Formal Concept Analysis: Foundations and Applications*, page 149–160, Berlin, Heidelberg. Springer-Verlag.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sarah Roscoe, Minal Khatri, Adam Voshall, Surinder Batra, Sukhwinder Kaur, and Jitender Deogun. 2022. Formal concept analysis applications in bioinformatics. *ACM Comput. Surv.*, 55(8).

Lütfi Kerem Şenel, Ihsan Utlu, Veysel Yücesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Stephen A. Vavasis. 2010. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377.

Rudolf Wille. 2004. Preconcept algebras and generalized double boolean algebras. In *Concept Lattices: Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23-26, 2004. Proceedings 2*, pages 1–13. Springer.

A Toy corpus

The corpus contains 24 synthetic sentences shown in Table 5. The target words—*king, queen, man, woman* and their plurals—are subjects of the sentences. Each of the eight words appears with three verbs—*live-in, wear, eat*—once for each. The remaining six words—*palace, house, tie, dress, alone, together*—discriminate the subject words so that they are in the analogical relations of three dimensions.

king live-in palace	*kings live-in palace*
queen live-in palace	*queens live-in palace*
man live-in house	*men live-in house*
woman live-in house	*women live-in house*
king wear tie	*kings wear tie*
queen wear dress	*queens wear dress*
man wear tie	*men wear tie*
woman wear dress	*women wear dress*
king eat alone	*kings eat together*
queen eat alone	*queens eat together*
man eat alone	*men eat together*
woman eat alone	*women eat together*

Table 5: 24 sentences in the toy corpus

B List of formal concepts

There are 28 formal concepts in the co-occurrence matrix derived from the toy corpus.

Suppose that the set of objects (target words) and the set of attributes (context words) be G, M respectively, defined as:

$$\begin{aligned} G = \{&king, man, queen, queens, \\ &kings, men, queens, women, \} \\ M = \{&tie, dress, \\ &palace, house, \\ &alone, together\} \end{aligned}$$

Then, all the formal concepts are identified as below:

$$\begin{aligned}
&T = (G, \emptyset) \\
&f_1 = (\{king, man, kings, men\}, \{tie\}) \\
&f_2 = (\{man, woman, men, women\}, \{house\}) \\
&f_3 = (\{king, queen, man, woman\}, \{alone\}) \\
&f_4 = (\{kings, queens, men, women\}, \{together\}) \\
&f_5 = (\{king, queen, kings, queens\}, \{palace\}) \\
&f_6 = (\{queen, woman, queens, women\}, \{dress\}) \\
&e_1 = (\{king, man\}, \{tie, alone\}) \\
&e_2 = (\{king, kings\}, \{tie, palace\}) \\
&e_3 = (\{man, men\}, \{tie, house\}) \\
&e_4 = (\{kings, men\}, \{tie, together\}) \\
&e_5 = (\{king, queen\}, \{palace, alone\}) \\
&e_6 = (\{man, woman\}, \{house, alone\}) \\
&e_7 = (\{kings, queens\}, \{palace, together\}) \\
&e_8 = (\{men, women\}, \{house, together\}) \\
&e_9 = (\{queen, woman\}, \{dress, alone\}) \\
&e_{10} = (\{queen, queens\}, \{palace, dress\}) \\
&e_{11} = (\{woman, women\}, \{house, dress\}) \\
&e_{12} = (\{queens, women\}, \{dress, togther\}) \\
&v_1 = (\{king\}, \{tie, palace, alone\}) \\
&v_2 = (\{man\}, \{tie, house, alone\}) \\
&v_3 = (\{kings\}, \{tie, palace, together\}) \\
&v_4 = (\{men\}, \{tie, house, together\}) \\
&v_5 = (\{queen\}, \{dress, palace, alone\}) \\
&v_6 = (\{woman\}, \{dress, house, alone\}) \\
&v_7 = (\{queens\}, \{dress, palace, together\}) \\
&v_8 = (\{women\}, \{dress, house, together\}) \\
&B = (\emptyset, M)
\end{aligned}$$

C Fuzzification of FCA

Formally, a fuzzy set A is a function $A : X \to L$ where X is a ground set and $L = [0, 1]$, which assigns the value to each member of X. A subsumption relation $A \subseteq B$ holds if and only if $A(x) \leq B(x)$ for all $x \in X$. In Fuzzy FCA, a formal concept is $\mathbb{K} := (G, M, I, L)$. We consider two fuzzy sets $A \in L^G, B \in L^M$ as objects and attributes and a fuzzy relation $I \in L^{G \times M}$. Mathematically, L can be generalized to a *residuated lattice* that includes $[0, 1]$ as its special case. Similar to the crisp setting, two fuzzy derivation operators $\uparrow\colon L^G \to L^M$ and $\downarrow\colon L^M \to L^G$ are defined as follows: For all $m \in M$ and $g \in G$,

$$A \uparrow (m) := \bigwedge_{g \in G} \big(A(g) \to I(g, m)\big) \in L \quad (12)$$

$$B \downarrow (g) := \bigwedge_{m \in M} \big(B(m) \to I(g, m)\big) \in L \quad (13)$$

Note that $A \uparrow \in L^M, B \downarrow \in L^G$ and $(\to) : L \times L \to L$, which is a binary operation defined on L. In plain English, the degree to which an object g belongs to the fuzzy set A should imply the level of co-occurrence between g and an attribute m, which retrospectively should determine the degree to which the attribute m belongs to another fuzzy set $A \uparrow$. Then, fuzzy formal concepts are defined as a pair of fuzzy sets (A, B) where $A \uparrow = B$ and $B \downarrow = A$ hold as in the crisp FCA.

We need to specify operations such as $(\to)$ to numerically compute them. Three specifications, named as Lukasiewicz, Gödel and Goguen, have already been proposed (Belohlavek and Vychodil, 2012), but instead we propose our own specification tailored to the analysis of a word co-occurrence matrix.

$$a \to b := \begin{cases} b/a & \text{if } a > 0 \\ \top & \text{if } a = 0 \end{cases} \quad (14)$$

where $\top$ is the greatest element in L. This specification is a slight modification of the one proposed by Goguen. The meet $\wedge$ is numerically calculated as a minimum.

Our specification is equivalent to defining $(A, A \uparrow)$ and $(B \downarrow, B)$ as a rank-one submatrix. Recall that the fuzzy set $A \in L^G$ assigns a value $x \in L$ to the element g. Similarly, the fuzzy set $A \uparrow \in L^M$ assigns a value $y \in L$ to the element m. Thus, the specification in Eq. (14) ensures that

$y = xy/x$ for $x > 0$. This means that nonnegative entries of both fuzzy sets $A, A\uparrow$ constitute a rank-one submatrix.

D Semantic categorization test

We used the four test sets for categorization test, Battig, BLESS, Series and Syntactic.

Battig test (Bullinaria and Levy, 2012), originated from Battig and Montague (1969), contains 53 categories with 10 words for each category, of which we used 44 categories in the experiments, since the others have less than two words in our vocabulary of the co-occurrence matrix.

BLESS (Baroni and Lenci, 2011) contains 17 categories with 5-17, of which we used 12 categories for the same reason.

Both of Series and Syntactic are developed by the authors to supplement Battig and BLESS, which contain only common nouns. Series is hinted by Hashimoto et al. (2016) that proposed the series completion task (*penny:nickel:dime:?*) for word embeddings. Syntactic is motivated by our early finding that comparative adjectives such as *quicker, faster, ...* emerge as a salient formal concept with a high threshold in the binary FCA experiment. In both test sets, each category consists of 4 to 5 words, which are manually selected by one of the authors. In the development process, we partly use AI assistance[5] to generate a list of candidates for a category and its word set, by prompting with an example "Direction: *north, east, south, west*".

Examples of a category in each test set are shown below (Table 6)

Test set	Category	Word set
Battig	Metal	*gold, iron, lead, steel,...*
BLESS	Fruit	*apple, banana, pear,...*
Series	Direction	*north, east, south, west*
Syntactic	Verb (go)	*go, goes, went, gone*

Table 6: Examples of test sets

We used only the categories that contain more than or equal to three words in our vocabulary, which are listed in Table 7.

E Decomposition by NMF

E.1 Decomposed submatrices by NMF

We applied NMF recursively in three rounds. In the first round, we decomposed the PPMI matrix as in $X_0 \approx W_1 H_1^T$ into 300 components ($\alpha = 0.0005$). In the second round, we applied NMF to the positive residual matrix after the first decomposition: $X_1 := \max(X_0 - W_1 H_1^T, 0)$ as decomposed as in $X_1 \approx W_2 H_2^T$ ($\alpha = 0.0003$). In the third round, the residual matrix $X_2 := \max(X_1 - W_2 H_2^T, 0)$ was decomposed into $X_2 \approx W_3 H_3^T$ ($\alpha = 0.0001$). Note that each component (rank-one matrix) $\boldsymbol{w}_k \boldsymbol{h}_k^T$ was forced to be sparse by L_1 regularization. Thus, their nonnegative rows and columns make a nonnegative rank-one submatrix, which we regard as a fuzzy formal concept.

The components derived in the first round were indexed from 1 to 300. Similarly, those in the second round were indexed from 301 to 600, and the ones from the third round were indexed from 601 to 900. We ordered each component by the Frobenius norm within each round. Therefore, the smaller ID number implies that the submatrix has a greater norm in each round.

Samples of the components are presented in Table 8. The class was evaluated by one of the authors according to the definition given in the Appendix E.2. The author also labeled a category from the words that comprise the submatrix $\boldsymbol{w}_k \boldsymbol{h}_k^T$. More specifically, for each vector $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$, we picked 20 words that correspond to the largest elements in the vectors, respectively. In Table 8, the only four top words are presented for both $\boldsymbol{w}_k$ as extents and $\boldsymbol{h}_k$ as intents. For ease of visibility, categories were labeled with more general expression, though they could be labeled with more focused category names.

Table 9 shows a supplemental analysis of the type of relatedness between words participating in each submatrix.

E.2 Types of qualitative classes

The set of words corresponding to the largest dimensions within each component is classified into four qualitative classes, as in the below definition (Table 10), following Lindh-Knuutila and Honkela (2015) . These classes indicate how well an identified formal concept (a rank-one matrix) is interpretable as a category.

[5] https://chat.openai.com/

Battig	**BLESS**	**Series**	**Syntactic**
Disease	Ground mammal	Emotion	Demonstrative adverb
Metal	Furniture	Season	Comparative adjective
Carpenter’s tool	Tool	Sea	Preposition
Crime	Container	Great lakes	Verb conjugation
Substance for flavoring food	Fruit	Direction	Manner adverb
Elective Office	Vehicle	Art form	Adverb of frequency
Toy	Appliance	Part of a tree	Personal pronoun
Weapon	Weapon	Book part	Linking verb
Member of clergy	Musical instrument	Continent	Demonstrative determiner
Four-footed animal	Building	Movie genre	Coordinating conjunction
Nonalcoholic beverages	Clothing	Number	Adjective of taste
Building for religious services	Bird	US president	Possessive pronoun
Precious stone		Stage of life	Frequency adverb
Part of human body		Planet	Quantitative determiner
Fruit		Weekday	Subordinating conjunction
Sport		Music genre	Action verb
Part of a building		Natural disaster	Modal auxiliary
Male’s first name		Decathlon	Total pronoun
Relative		Family	Adjective of size
Human dwelling		Ocean	Interrogative pronoun
Insect		Adverb of time	Article
Type of fuel		Month	Totality adverb
Music instrument		Communication act	Verb conjugation
Furniture		Match	
Ship		Religion	
Kind of money		Time of day	
Color		Writing	
Kind of cloth		Style of architecture	
Unit of distance		Midwest U.S. state	
Type of music			
City			
Country			
Reading material			
Military title			
Natural earth formation			
Unit of time			
Part of speech			
Kitchen utensil			
Vehicle			
Science			
Weather phenomenon			
Occupation or profession			
Bird			

Table 7: Used categories of the test sets

ID	**Class**	**Category**	**Extents** (top 4 words)	**Intent** (top 4 words)
2	D	Geography	*iran, kerman, khorasan, province*	*iran, kerman, khorasan, province*
5	N	None	*pineapples, tasteful, lilongwe, unimpressive*	*dawn, windsor, batting, relegation*
8	D	Music	*chart, charts, billboard, singles*	*chart, charts, billboard, singles*
14	D	Sports	*discus, javelin, jump, hurdles*	*discus, javelin, jump, hurdles*
22	D	Education	*degree, bachelor, doctorate, laude*	*degree, bachelor, doctorate, laude*
35	D	Diplomacy	*embassy, ambassador, diplomatic, relations*	*turkmenistan, tajikistan, kyrgyzstan, uzbekistan*
46	D	Sports	*baseman, pitcher, outfielder, shortstop*	*baseman, pitcher, outfielder, shortstop*
89	D	Religion	*rabbi, yeshiva, synagogue, hebrew*	*rabbi, yeshiva, synagogue, hebrew*
90	D	US states	*idaho, montana, dakota, wyoming*	*idaho, montana, dakota, wyoming*
95	D	Climates	*cyclone, hurricane, storm, typhoon*	*cyclone, hurricane, storm, typhoon*
98	D	Politics	*polling, votes, voters, vote*	*polling, votes, voters, vote*
102	D	Phrases	*increases, decreases, decrease, increase*	*temperature, concentrations, accuracy, velocity*
104	D	Politics	*incumbent, reelection, democrat, republican*	*incumbent, reelection, democrat, republican*
116	P	Medical	*ligament, knee, ankle, injury*	*ligament, knee, ankle, injury*
125	P	Career	*postdoctoral, professor, adjunct, emeritus*	*postdoctoral, professor, adjunct, emeritus*
137	P	TV show	*starring, roommate, daughters, actress*	*jennifer, laura, jessica, nicole*
146	P	Legal	*convicted, guilty, sentenced, imprisonment*	*convicted, guilty, sentenced, imprisonment*
147	P	History	*nazi, nazis, deported, camps*	*nazi, nazis, deported, camps*
159	P	Geography	*mountain, peaks, summit, mountains*	*mountain, peaks, summit, mountains*
160	M	Expression	*acclaim, garnered, reviews critical*	*garnered, acclaim, reviews, critical*
165	P	Expression	*regain, recover, conquer, attract*	*trying, attempting, attempt, attempts*
181	M	Expression	*tasked, thereby, prevented, intention*	*securing, obtaining, capturing, creating*
184	M	Expression	*lied, intentions, poisoned, whereabouts*	*reveals, realizes, believing, realises*
192	D	Music	*punk, hop, hip, folk*	*punk, hop, hip, folk*
210	D	Religion	*christianity, catholicism, islam, beliefs*	*christianity, catholicism, islam, beliefs*
212	M	Religion	*you, think, really, know*	*you, think, really, know*
214	M	Syntactic	*various, numerous, several, these*	*genera, disciplines, locations, dialects*
237	D	Comparative	*faster, stronger, heavier, than*	*faster, stronger, heavier, than*
239	D	Politics	*obama, barack, reagan, clinton*	*obama, barack, reagan, clinton*
313	D	Religion	*quantities, amounts, sums, amassed*	*enormous, huge, immense, considerable*
329	P	Time	*spends, spend, spent, spending*	*summers, much, time, remainder*
341	P	Geography	*maui, oahu, hawaii, honolulu*	*maui, oahu, hawaii, honolulu*
370	D	Unit	*millions, billions, million, billion*	*millions, billions, million, dollars*
405	M	Linguistics	*vowel, vowels, stressed, accent*	*vowel, vowels, stressed, accent*
408	P	Travel	*immigration, nationals, emigration, citizen*	*immigration, nationals, emigration, citizen*
419	D	Proposal	*proposal, offer, invitation, plea*	*rejected, accepted, rejects, accepting*
431	M	Expression	*poorly, properly, carefully, fully*	*handled, treated, understood, trained*
435	D	Buildings	*housed, built, constructed, build*	*synagogue, mosque, mansion, convent*
484	D	Auxiliary	*did, does, doesn, didn*	*speak, exist, suffer, appear*
507	D	Movement	*down, forth, out, into*	*fell, put, falling, fallen*
514	D	War	*pistol, revolver, magnum, rifle*	*pistol, revolver, magnum, rifle*
517	M	Plants	*botanical, zoological, garden, gardens*	*botanical, zoological, garden, gardens*
577	P	Expression	*totally, completely, virtually, almost*	*totally, virtually, completely, vanished*
605	D	Number	*vii, ix, viii, xiii*	*fantasy, corps, intensity, chapter*
626	P	Accounting	*collect, collecting, exception, collected*	*taxes, debt, debts, fees*
645	D	Month	*june, july, august, september*	*premiered, consecrated, baptised, inaugurated*
667	D	Expression	*taking, take, taken, takes*	*hostage, advantage, seriously, refuge*
669	D	Geography	*gaza, palestinians, palestinian, israeli*	*strip, gaza, rockets, barrier*
679	D	Geography	*colombian, venezuelan, peruvian, chilean*	*peso, divisi, primera, aut*
774	P	IT	*java, server, windows, software*	*java, server, windows, software*
781	D	Expression	*bought, purchased, buying, buy*	*shares, stake, tickets, tracts*
784	M	Marketing	*advertising, commercials, campaigns, marketing*	*advertising, commercials, campaigns, marketing*
804	M	Expression	*about, detail, matters, topics*	*discuss, discussed, discussing, discusses*
855	M	Expression	*heavily, originally, by, recently*	*influenced, inspired, invented, borrowed*
864	D	Expression	*currently, presently, still, today*	*currently, resides, owns, produces*
874	P	Expression	*launching, pursued, launched, developed*	*ventures, venture, scheme, initiative*

Table 8: Samples of decomposed submatrices labeled with a category name. Classes are abbreviated; D:Descriptive, P:Partial, M:Meaningful, N:Nonsense

Proximity	**R1**	**R2**	**R3**
Categorical	74	64	53
Contextual	171	147	148
Combinatorial	41	59	62
Syntactic	9	18	19
None	5	12	18
Total	300	300	300

Table 9: Proximity types of word relations in each NMF-decomposed component. Categorical: words are in the same category, Contextual: words are related in a shared context, Combinatorial: words are a part of possible phrases, i.e., paradigmatic, Syntactic: words are in the same syntactic category.

Class	**Description**
Descriptive	Words are related in some way, and the majority label given is as descriptive as possible of the words in the set.
Partial	Words are related in some way, and the majority label is somewhat descriptive, but a more descriptive account can be easily given.
Meaningful	Words are related, but no majority label describes the words.
Nonsense	There is no majority label, nor is there any perceived relation between the words in the set.

Table 10: Definition of qualitative classes assessing how well the labels describe the words in each formal concept. (Lindh-Knuutila and Honkela, 2015)