Design Science

Knowledge of technological artifacts: Toward linguistic and structural foundations from patent descriptions

L. Siddharth[1] and Jianxi Luo[2]

[1]*Engineering Management and Systems Engineering, The George Washington University School of Engineering and Applied Science, USA*
[2]*Department of Systems Engineering, City University of Hong Kong, Hong Kong*

Abstract

Design and innovation processes primarily synthesize the knowledge of existing technological artifacts. Understanding the foundations of such artifact-level knowledge is essential for enabling the knowledge retrieval and representation that govern these syntheses. In this study, we analyze a large, stratified sample of 33,881 patent descriptions across the total technology space. We populate knowledge graphs of these descriptions by combining factual triplets (entity:: relationship:: entity) extracted at the sentence level. From these knowledge graphs, we uncover the linguistic and structural foundations of the knowledge of technological artifacts. Linguistically, we identify syntactic patterns that explain how entities and relationships are constructed at the term level. Structurally, we identify motifs, including dominant 3-node and 4-node subgraph patterns, that reveal how entities and relationships are combined locally in artifact descriptions. Delving into these motifs reveals that natural language artifact descriptions primarily capture the design hierarchy of artifacts. At a local level within artifact descriptions, the motif analyses reveal that only abstract technical knowledge is captured, indicating potential limitations of relying on text-mining for knowledge-intensive tasks. Based on these observations, we propose and demonstrate knowledge specification strategies that can help simplify and modularize knowledge structures populated from technological artifact descriptions.

Keywords: Technological artifacts, Patent descriptions, Design knowledge, Knowledge graphs, Computational linguistics, Network motifs

Received 03 August 2025
Revised 05 May 2026
Accepted 06 May 2026

Corresponding author:
Jianxi Luo
jianxi.luo@cityu.edu.hk

Des. Sci., vol. 12, e17
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2026.10064

the Design Society a worldwide community

Cambridge University Press

1. Background

Design and innovation processes fundamentally involve generating knowledge through the retrieval and synthesis of the knowledge of existing technological artifacts (Liu, Tao, & Bi 2022). Ontologies explicitly specify how knowledge is conceptualized in a domain of discourse (Gruber 1993; Chandrasegaran *et al.* 2013). They have been central to engineering design research in knowledge retrieval and representation, where scholars theorize and operationalize the basis of knowledge represented, retrieved and reused during design and innovation processes (Siddharth, Blessing, & Luo 2022a). Their contributions stem from different perspectives, such as:

- The *systems engineering perspective* (Simon 1962) suggests that artifacts constitute various parts that interact in different ways – largely involving material, energy and information exchanges via shared interfaces (Browning 2001) – such that the whole is more than the sum of its parts.

- The *function structure perspective* (Rodenacker & Rodenacker 1976) simplifies the overall function of a technological artifact into sub-functions that are connected using flows, represented using a prescribed vocabulary (Hirtz *et al.* 2002) for both functions (for example, import and transfer) and flows (for example, liquid and signal).
- The *qualitative physics perspective* (De Kleer & Brown 1984) describes an artifact using constructs such as structure (what?), behavior (how?) and purpose (why?), leading to various theories and models on functional representation (Chandrasekaran & Josephson 2000).
- The Core Product Model (Fenves *et al.* 2008) developed at the National Institute of Standards and Technology (NIST) integrates these perspectives into a scheme having object (for example, geometry, material) and relationship (for example, constraint, part-of) classes.

In technology and innovation management fields as well, scholars theorize the foundations of the knowledge of technological artifacts. Herein, scholars study the knowledge of technological domains to understand the mechanisms of technological change. They capture domain-level knowledge using networks of topics, co-inventors, co-classifications, co-assignees and co-citations, predominantly populated from patent databases (Siddharth, Li, & Luo 2022b; Li, Siddharth, & Luo 2023). For instance, Malhotra *et al.* (2021) assess the citation network of 101,260 patents under lithium-ion batteries and find 491 patents forming the core of the network and governing the knowledge trajectory.

These perspectives only reflect expert views on how the knowledge of an artifact should be systematically captured. Robinson *et al.* (2023) studied the sewing machine patents granted in the last 200 years to model the “dominant” functional architecture using techniques proposed in the functional basis (Hirtz *et al.* 2002), leveraging an expert-defined ontological lens to study the domain of interest. In addition, the studies in technology management seem to rely on meta-level knowledge given by citations, inventors, classifications and others, which themselves represent an expert-defined scheme of technology space, without considering the actual knowledge of artifacts documented in vast text documents (Siddharth *et al.*, 2022a, 2021). Considering these broad research gaps, we address the following research question from a different perspective.

What is the basis of the knowledge of technological artifacts? To address the above, we procure the knowledge of technological artifacts from a large sample of patent descriptions. Specifically, we populate knowledge graphs of patent descriptions using the sentence-level fact extraction methods developed in our prior work (Siddharth & Luo 2024; Siddharth 2024a). Using the knowledge graphs, we implement techniques from computational linguistics and networks to reveal the linguistic and structural foundations of the knowledge of technological artifacts. By leveraging patent descriptions (not the metadata) and bottom-up (not expert-defined), sentence-level fact extraction for populating knowledge graphs, our study addresses the overall research gaps observed in both engineering design and technology management literature.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

2. Data and methods

2.1. Knowledge graph data

Patent documents offer rich and detailed descriptions of several million technological artifacts spanning the total technology space (Siddharth, Li, & Luo 2022b). Despite this advantage, patent documents remain underutilized due to the lack of methods to explicitly capture the knowledge structures of artifacts representing them. To address this limitation, we recently developed a method (Siddharth & Luo 2024; Siddharth 2024a) to extract sentence-level facts of the form entity:: relationship:: entity from patent documents. The following is an example of facts extracted from a sentence taken from a glue gun patent[1].

> "The adhesive, softened by the heating step, can be readily extruded mechanically by means of a piston from the front end of the glue gun."
>
> The adhesive:: softened by:: the heating step
> The adhesive:: readily extruded mechanically by:: means
> means:: of:: a piston
> means:: from:: the front end
> a piston:: from:: the front end
> the front end:: of:: the glue gun

Exhibiting up to 99.7% accuracy, the method for extracting facts – as illustrated above – relies on a two-stage transformer-based language model trained on a proprietary dataset of 50,000 patent sentences and facts[2]. Applying this method sentence-wise to a patent description results in a list of facts that can be combined into a knowledge graph, as illustrated in Figure 1. We utilize our prior method to populate knowledge graphs – illustrated further in Figure 2 – of a large sample of patent descriptions and utilize these knowledge graphs to study the foundations of the knowledge of technological artifacts. To gather the sample of patents, we leverage Patents View[3] and source patent information as of 8 June 2023, by which the database included over 8.2 million granted US patents. We restrict our scope to utility patents that are granted for their functionalities rather than aesthetics or other criteria.

Considering the 7.3 million utility patents in the population, we use a Qualtrics calculator[4] to determine a reasonable sample size of 16,540 according to a 99% confidence level and 1% margin of error. The calculator leverages the Cochran's formula (Ahmed 2024, p. 5) as shown in Equations 1 and 2.

$$N_o = \frac{Z^2 p(1-p)}{E^2} \tag{1}$$

$$N_s = \frac{N_o}{1 + \frac{N_o}{N_P}} \tag{2}$$

Where N_o is the initial sample size calculated using the Z-score (Z), Error (E) and p-value (p). Given the 1% margin of error, E is 0.01 – Z is assigned 2.576 corresponding to 99% confidence value and p is assigned 0.5 as the default. The resulting N_o is adjusted for our population size, N_P = 7,373,519 (see Table 1), as in Equation 2. The final sample size, as per the calculation above, is returned as 16,540, subject to numerical approximations in the Qualtrics calculator. We

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

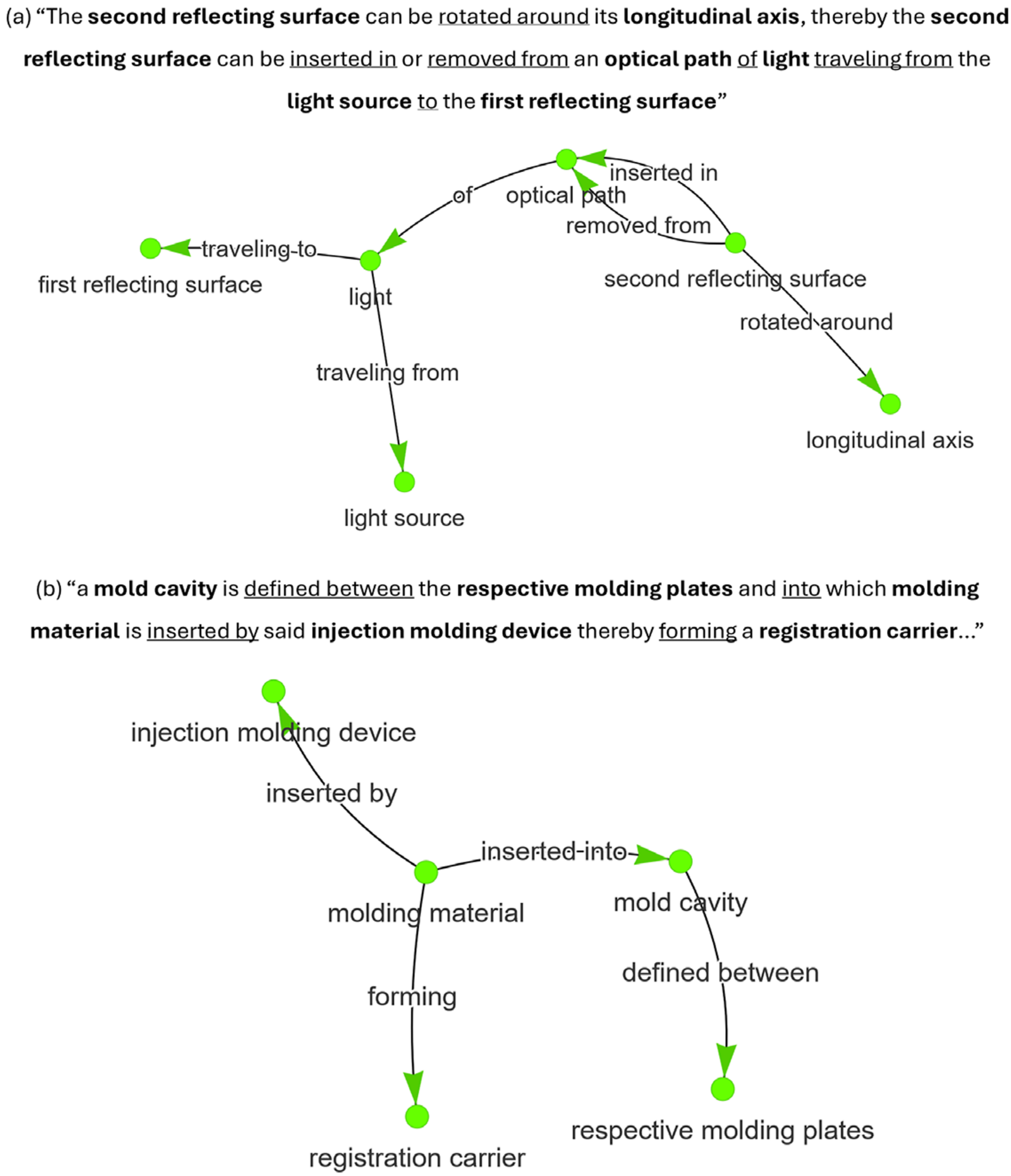

Figure 1. Illustrating explication of knowledge using a sentence from patented artifact descriptions of (a) vehicle light (Vehicle light with movable reflector portion and shutter portion for selectively switching an illuminated area of light incident on a predetermined portion of the vehicle light during driving – https://patents.google.com/patent/US6796696/) and (b) injection molding (Installation for manufacturing registration carriers – https://patents.google.com/patent/US5451155/).

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

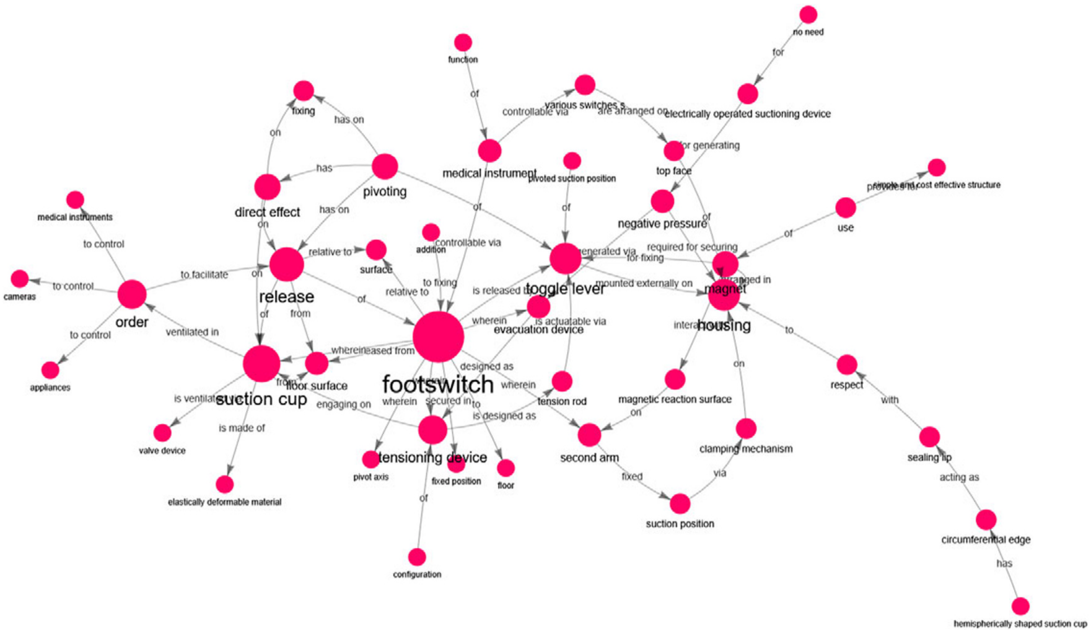

Figure 2. Example knowledge graph extracted from a patent (Footswitch for a medical instrument – https://patents.google.com/patent/US10437277B2/) using our prior method (Siddharth & Luo 2024). Each of the 33,881 patents in the sample data has a knowledge graph as illustrated here.

Table 1. Sampling patents and extracting design knowledge from USPTO

Sampling patented artifact descriptions	
Total number of granted patents	8,260,142
Upon filtering WIPO kind to A, B1, B2	7,484,623
Upon selecting utility patents	7,484,622
Upon filtering a minimum of one claim	7,373,519
Estimated sample size	16,540
Upon sampling	33,945
Upon scraping the full text	33,884
Extracting design knowledge	
Number of patents with facts	33,881
Number of sentences	7,191,733
Number of sentences with facts	5,451,709
Number of facts	24,537,587
Number of unique entities	5,015,681
Number of unique relationships	845,303

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

intended the sample to be stratified according to patent counts as per the CPC 4-digit classification scheme[5] (for example, A02F). By constraining the distribution such that there is no overlap among classes, we obtain a sample of 33,945 patents (Table 1). For the sample of 33,945 patents, full text is available for 33,884 patents, which we clean and split into sentences[6]. For each sentence in each patent, we apply our method (Siddharth & Luo 2024; Siddharth 2024a) to extract facts of the form head entity:: relationship:: tail entity. When these facts are combined across sentences and within a patent, a knowledge graph can be obtained, as illustrated in Figure 2. The knowledge graph data populated[7] for this work, including patent information, knowledge graphs and analysis, is publicly available (Siddharth 2024b).

2.2. Linguistic foundations – Zipf distribution analysis

By studying the entities and relationships that form the knowledge graphs (for example, Figure 2), we uncover the linguistic foundations of the knowledge of technological artifacts. In Table 1, we noted that the 33,881 patent knowledge graphs together comprise 5,015,681 unique entities and 845,303 unique relationships. To understand how these entities and relationships are linguistically constructed, we transform them into syntaxes[8], for example, “a shake” → “a NN,” “the cured spar assembly” → “the JJ NNP NNP,” “all three erase blocks” → “all CD NN NNS,” where the parts-of-speech NN = Noun, JJ = adjective, *CD* = cardinal number and others. In Appendix A, we provide the full list of parts of speech, their definitions and examples.

Upon transforming the terms that constitute patent knowledge graphs, we obtain 408,323 unique entity syntaxes and 73,352 unique relationship syntaxes. We analyze the frequencies of the unique syntaxes to find the most generalizable ones that form the linguistic foundations of the knowledge of technological artifacts. In computational linguistics, term frequencies in a corpus are typically studied using the Zipf distribution (Mollica & Piantadosi 2019). In engineering design literature, Murphy *et al.* (2014) analyze 65,000 patent abstracts using Zipf distribution to identify the most representative 1,700 functions (verbs) of patented artifacts. Zipf distribution follows the probability mass function (PMF) as shown in Equations 3 and 4.

$$PMF(k;s,N) = \frac{1}{H_{N,s}} \cdot \frac{1}{k^s} \tag{3}$$

$$H_{N,s} = \sum_{k=1}^{N} \frac{1}{k^s} \tag{4}$$

In the above equations, N is the number of distinct items, k stands for the rank of an item in the frequency distribution and s is the distribution parameter. The idea behind the Zipf distribution is that the probability of an item in the corpus is inversely related to its rank – k raised to the power – s. To ensure the probabilities of all terms sum to 1, a generalized harmonic number ($H_{N,s}$), as in Equation 4, is multiplied by the PMF. Herein, we assume N to be finite – in cases where $N \to \infty$, $s > 1$. Based on the frequency proportions of entity and relationship syntaxes, we perform a curve fitting using SciPy[9] on Equation 3 and examine the fitted distribution and cumulative distribution plots to view the syntaxes at different percentiles.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

2.3. Structural foundations – motif analysis

By studying the network structures of the knowledge graphs (for example, Figure 2), we uncover the structural foundations of the knowledge of technological artifacts. The structures of patent knowledge graphs have building blocks, referred to as motifs, that can be identified as dominant 3-node and 4-node subgraph patterns. Motif discovery has been popular in studies on sensory transcription/transduction networks and ecological food webs (Milo *et al.* 2002; Alon 2006; Strona *et al.* 2014). We leverage the approaches from these studies and adapt them to patent knowledge graphs. In engineering design literature, Fu *et al.* (2013) use latent semantic analysis (LSA) to extract functional and surface-level similarity between a set of 100 patents. They implement a Bayesian structural-form algorithm to search candidate forms such as trees, rings and hierarchies that explain the similarity patterns.

2.3.1. Subgraph mining

The first step is to mine 3-node and 4-node subgraphs. Since this task is NP-hard, it is usually implemented with specialized algorithms and/or restricted to specific patterns (Ribeiro *et al.*, 2021). However, in our work, we implement a simple technique to mine 3-node and 4-node subgraphs. If a graph includes nodes 1, 2, 3, 4…, we filter edge pairs that have three unique nodes, for example, edges 1–2 and 2–3, and discard the pairs that have two or four unique nodes, for example, edges 1–2 and 3–4, 2–3 and 3–2. Upon populating triples that represent 3-node subgraphs, we pair triples with edges and filter those that have four unique nodes, for example, 1–3–2, 2–4. This approach is not suitable for densely connected networks. However, the graphs in our data are quite sparse; for example, the largest graph[10] has 4460 nodes and 8204 edges, indicating a mean degree = 1.839.

2.3.2. Pattern matching

The second step is to categorize the subgraphs into unique patterns. Each subgraph belongs to a unique pattern (for example, Figure 3) that may not be directly identifiable because subgraphs can be visually oriented in different ways (also called isomorphs) and the nodes/edges can be labelled differently. To match the subgraphs into unique pattern categories, it is necessary to extract unique features and represent them using canonical forms (Chomsky 1965). As shown in Figure 3, we find and sort the in- and out-degrees of nodes and the number of edges between individual node pairs. The numeric identifier stated alongside the examples shown in Figure 3 can be considered canonical forms that are identical for all isomorphs of a subgraph pattern. We find exactly 13 distinct canonical forms for 3-node subgraphs, consistent with 13 patterns in the literature (Milo *et al.* 2002, p. 824). In total, we identify 13 and 134 patterns[11] of, respectively, 3-node and 4-node subgraphs, as shown in Appendix B.

2.3.3. Randomizing graphs

The third step is to randomize each patent knowledge graph to create a null model that could help in understanding the significance of a subgraph pattern in the original graph. Such randomization should retain some properties of the original graph. In biology literature, (Alon 2006, p. 27) generates an ensemble of

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

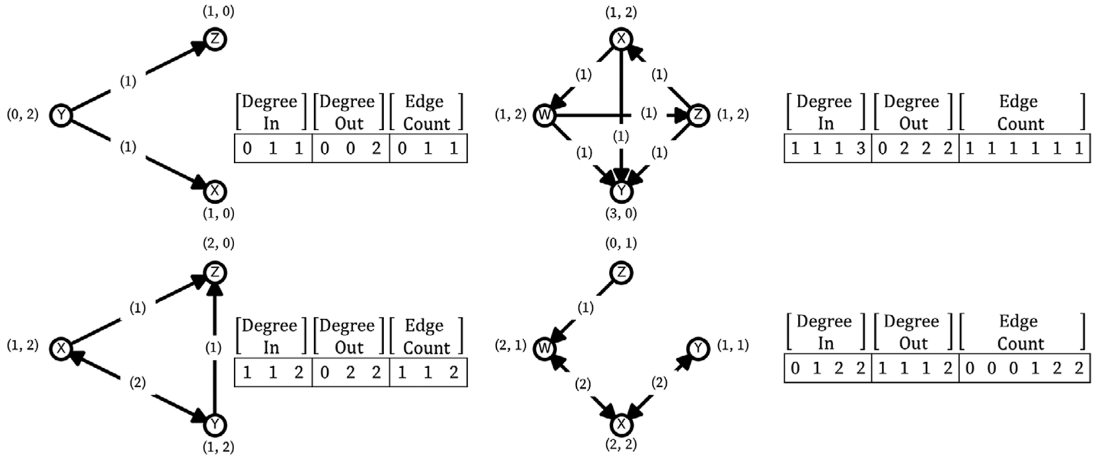

Figure 3. Obtaining canonical forms for 3-node and 4-node patterns to match with isomorphs.

Erdos–Renyi random graphs with the same number of nodes as that of sensory–transcription networks. Watts & Strogatz (1998, pp. 440, 441) propose rewiring of edges that induce randomness, preserving the number of nodes as well as edges. In ecological networks, Stone, Simberloff, & Artzy-Randrup (2019, p. 4) propose dual edge swapping between edges such that the degree distribution is also retained. As edge swapping can take up to 50,000 iterations to attain a 50% perturbed random graph, Strona *et al.* (2014, pp. 3, 4) propose a computationally efficient "curveball" algorithm, which we adopt in our work. We prioritize computational efficiency primarily because of the sparsity of the patent knowledge graphs (mean degree = 1.839) that require several iterations to output a distinguishable randomized graph – preferably with 50% different edges. As illustrated in Figure 4, edge swapping is done between two randomly selected nodes until 50% perturbation is achieved. While swapping, we additionally include a constraint that a node cannot self-relate,[12] which was not imposed in Strona *et al.* (2014).

2.3.4. Analyzing differences

The final step involves analyzing the differences in the number of 3-node and 4-node subgraphs for each pattern between the original and randomized versions for all patent knowledge graphs. If a pattern occurs significantly more frequently in the original graph compared to the randomized version, the pattern can be considered a motif. For each graph representing a patent i, let us consider that a subgraph pattern p occurs X_{ip} and Y_{ip} times in the original and randomized versions, respectively. We normalize the difference between these counts using the edge count (E_i) as in Equation 5 and measure the z-score of this normalized difference for all graphs in the dataset as in Equation 6, where mean (μ_p) and standard deviation (σ_p) are measured across all 33,881 patent knowledge graphs.

$$\Delta_{ip} = \frac{X_{ip} - Y_{ip}}{E_i} \tag{5}$$

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

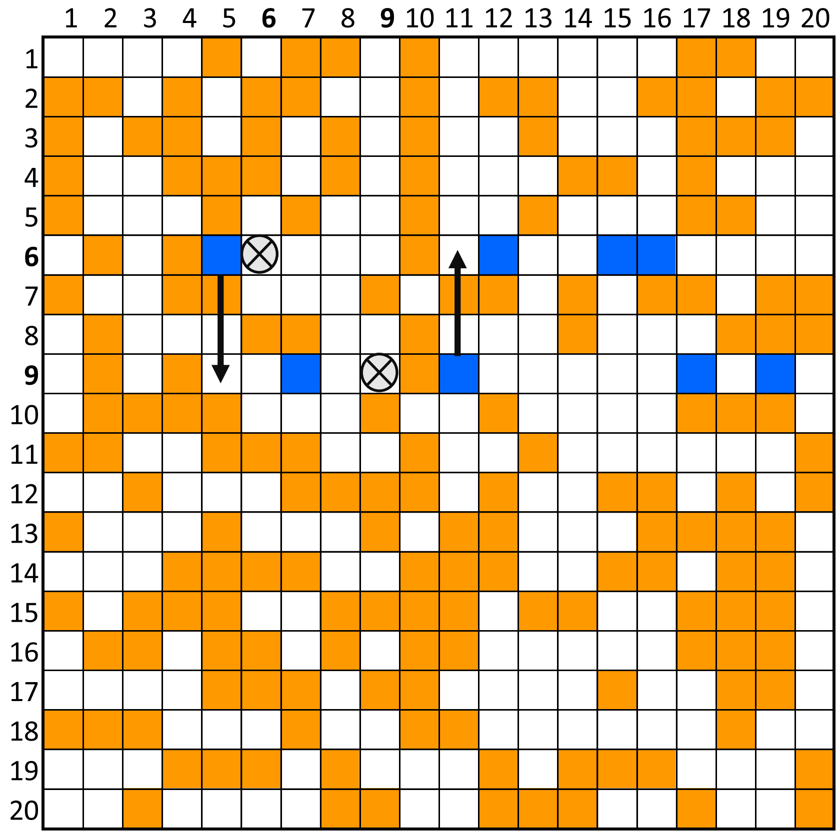

Figure 4. Illustrating the curveball algorithm.

$$Z_{ip} = \frac{\Delta_{ip} - \mu_p}{\sigma_p} \tag{6}$$

The z-score (Z_{ip}) as calculated in Equation 6 could tell us whether a pattern p observed in a patent graph i is significantly higher than it would be by chance. At 95% confidence, we cap the z-score to 1.64 to tell whether a pattern p is a motif of the graph i. We repeatedly perform this calculation for each 3-node and 4-node pattern (listed in Appendix B) in each patent knowledge graph.

3. Results and discussion

3.1. Linguistic foundations – entity and relationship syntaxes

Following the methods described in Section 2.2, we transform the entities and relationships in patent knowledge graphs into unique syntaxes and rank them according to their frequencies. Upon fitting the frequency proportions of entity and relationship syntaxes[13] (Figure 5a,b) onto the probability function as in Equation 3, we discuss the syntaxes at different percentiles using the cumulative distribution as in Figure 5c,d.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

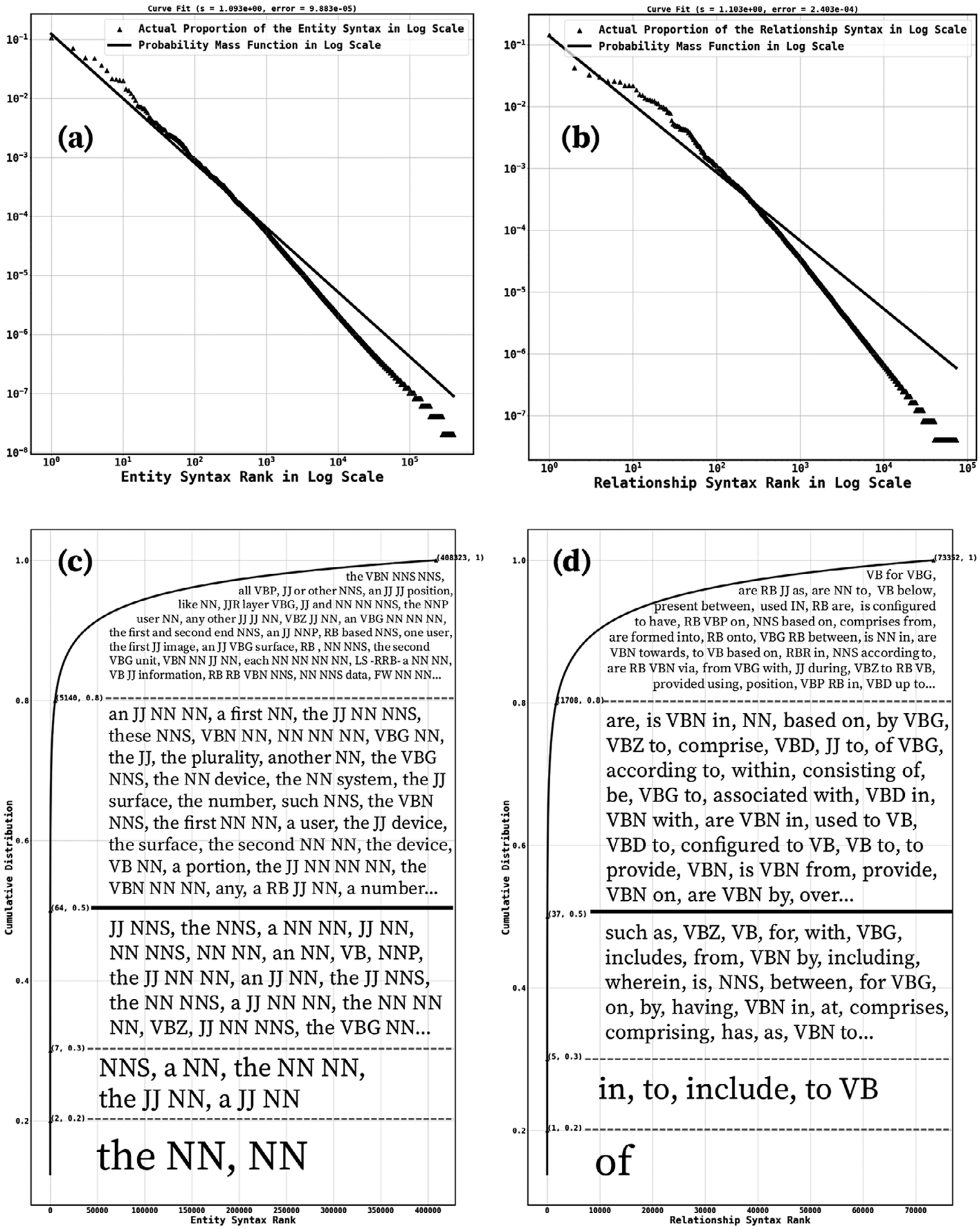

Figure 5. Probability Zipf distribution and cumulative Zipf distribution of (5a and 5c) entity and (5b and 5d) relationship syntaxes. Syntaxes are linguistic forms expressed using frequent words and parts of speech tags like NN, VB, JJ and others.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

In the first half of the entity syntax distribution (Figure 5c), single nouns (NN, the NN, for example, cardboard, the processor) constitute the most basic and frequent linguistic forms, which are a commonly recognized form for identifying components, processes and all constituents of an artifact in engineering design literature (Hirtz *et al.* 2002; Fu *et al.* 2013; Fantoni *et al.*, 2021). As single nouns cannot often pack a lot of information, these are extended to relatively complex forms with other nouns (proper – NNP and plural – NNS), adjectives (JJ) and verbs (VB). In the second half of the entity syntax distribution, we can observe several complex forms, which, according to Hagoort (2019), are formed by hierarchically associating basic forms, for example, "the front facing surface" = ["facing" ["front" ["surface"]]]. Hence, the most basic forms, that is, 64 syntaxes in the first half of the distribution, constitute the linguistic foundations for complex entity syntaxes and, in general, the knowledge of technological artifacts.

In the first half of the relationship syntax distribution (Figure 5d), "of" constitutes the most basic and frequent relationship, commonly used to capture the attribute, subsystem or subprocess of a larger entity. Besides, "in" and "to" capture various structural and behavioral relationships depending on the context. For instance, "to" can be mentioned in the transition sense, for example, "changed from… to…," in an introductory sense, for example, "related to… and to…," or in a utility sense, for example, "to deliver…" as expressed by the "to VB" syntax. Along with these, we observe "such as" – an exemplar relationship that is often used in instances like "volatile components such as X, Y, Z…." We then observe single verbs ("VBZ," "VB"), which are considered to be common linguistic representatives of relationships (Jamrozik & Gentner 2020; Siddharth *et al.*, 2021). The hierarchical relationship "include" is the most important part of an artifact description that provides a systemic view and helps integrate and explain other constituents. To understand the hierarchical relationship further, we capture all forms of hierarchical relationships by measuring semantic similarity with "include" using Sentence Transformers[14] and display their generalizable forms in Figure 6.

3.2. Structural foundations – motifs

Following the methods described in Section 2.3, we identify the motifs for each patent knowledge graph in the dataset of 33,881 knowledge graphs[15]. Across patents, we select the motifs that occur significantly higher than the others with 99% confidence[16]. Such motifs can be considered to constitute the structural foundations of the knowledge of technological artifacts. As shown in Figure 7, we select such motifs that are dominant overall or within the largest domains[17] given by the CPC scheme. These motifs have also been identified in previous studies. We recall them to know whether their interpretations can help us understand the motifs that we have identified from the knowledge graphs of patent descriptions.

In patent citation networks, Zhang, Zhang, & Wang (2024) mention the recurrence of direct citation (or sequence) and co-citation (or aggregation) as represented by Patterns 11 and 13, respectively. They note that such triangular structures induce an "echo-chamber effect" in which such structures reinforce themselves, leading to greater connectivity and lower entropy. Hence, Patterns 11 and 13 should contribute greatly to the connectivity of overall knowledge graphs

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

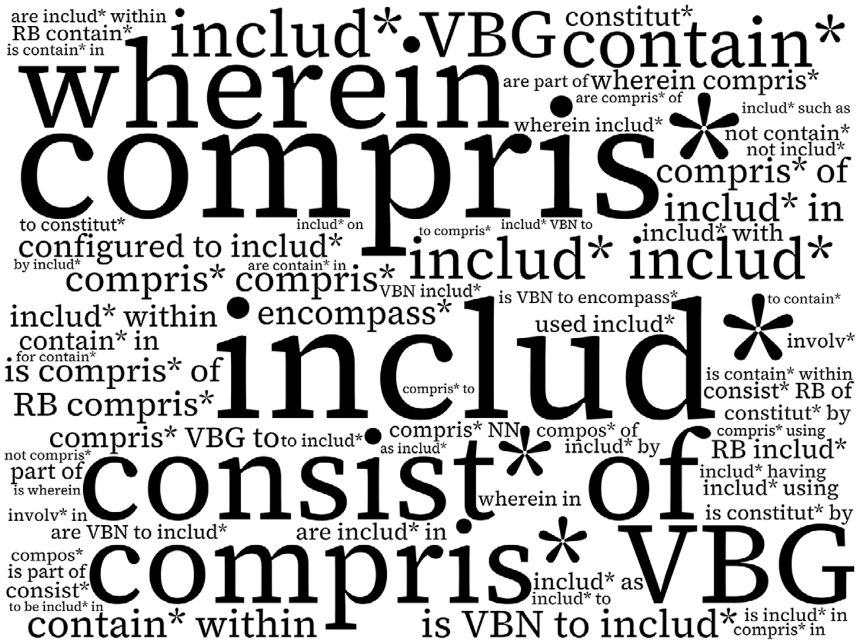

Figure 6. Seventy-three hierarchical relationship syntaxes.

of patents and thus form an imperative part of the structural foundations of the knowledge of technological artifacts. Pattern 11 is termed "sequence" in Fu *et al.* (2013, p. 5), where representative structures in patent–patent similarity networks were identified. Patterns 130 and 141 include a parallel link between two nodes that are sequentially connected. Fu *et al.* (2013, p. 5) refer to such representations as "order," whereas Alon (2006, p. 45) labels them as "feed-forward loop" as found in sensory transcription networks.

Pattern 8 resembles "within-unit reciprocity," which is found in organizational interaction networks (Brennecke *et al.* 2021, p. 8). Surprisingly, none of the motifs represent or extend "cyclic-closure" [↻], which is a common conceptualization scheme in various fields (Brennecke *et al.* 2021, p. 8), while also being identified among the patent structures in Fu *et al.* (2013, p. 5). Patterns 122 and 125 structurally resemble "bi-fan," which is a common motif in transcription networks (Alon 2006, p. 93). These patterns also accommodate hierarchy [$\leftarrow \cdot \rightarrow$], which is individually not a dominant motif from our observations, despite being commonly used for visualizing nested hierarchies and process flows (Siddharth, Chakrabarti, & Ranganath 2020) and also identified among the representative patent database structures in Fu *et al.* (2013, p. 5).

While existing studies offer unique interpretations for different motifs, their terminologies, "bi-fan," "feed-forward loop" are specific to those networks where an edge carries a specific meaning. Such terminologies only provide limited interpretive capabilities for the motifs that we have identified from patent knowledge graphs. In Figures 8 and 9, we present the qualitative subgraphs constituting the motifs identified and showcased in Figure 7. As the motifs alone do not offer

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

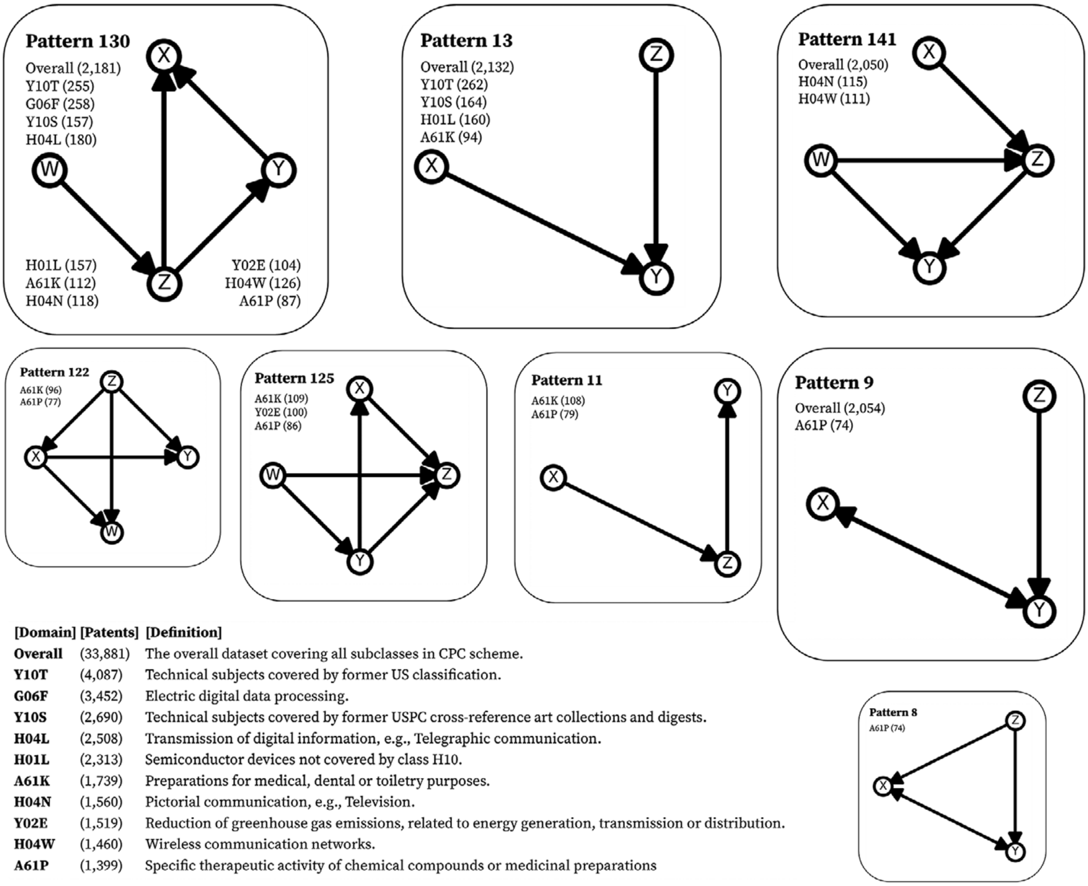

Figure 7. Dominant motifs overall and within the largest classes. For each motif pattern, we indicate the number of patents in "(.)" alongside the domain code.

sufficient interpretive capabilities, we populate the subgraphs with edge labels falling under each subgraph pattern identified as motifs in Figures 8 and 9.

While the subgraphs in Figures 8 and 9 offer relatively better interpretation than pure motifs, we examine their frequencies to determine their generalizability. For example, the most frequent subgraph under Pattern 125 has a frequency of 483, which constitutes a minor portion (0.14%) of the raw motif count = 345,797, while the top three patterns together constitute merely 0.375% of the raw count. Such frequencies indicate that the interpretations of the qualitative subgraphs within the motifs are not sufficiently generalizable, even though the motifs (pure structure) are generalizable constituents of knowledge structures of technological artifacts. Among the patterns shown in Figures 8 and 9, we observe that the most frequent subgraphs under Patterns 122, 13 and 11 have frequencies large enough to provide relatively more generalizable interpretations compared to the remaining patterns[18]. For example, the most frequent subgraph in Pattern 122 has a frequency proportion of 25.499%, while the top three subgraphs together constitute 32.426%. Hence, the interpretations made using the most frequent subgraphs under

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

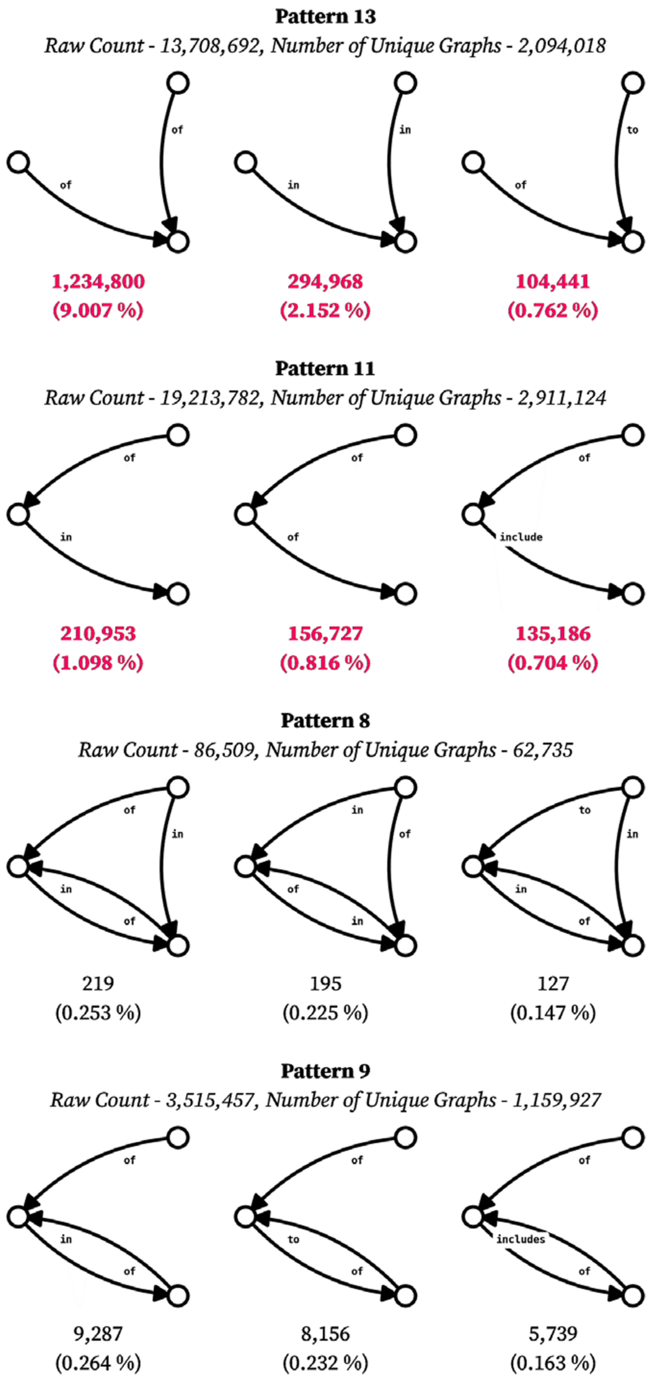

Figure 8. Top three most frequent subgraphs under the motifs represented by Patterns 13, 11, 8 and 9. The frequency of each subgraph and its percentage with respect to raw motif count are mentioned in the (.) beneath.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

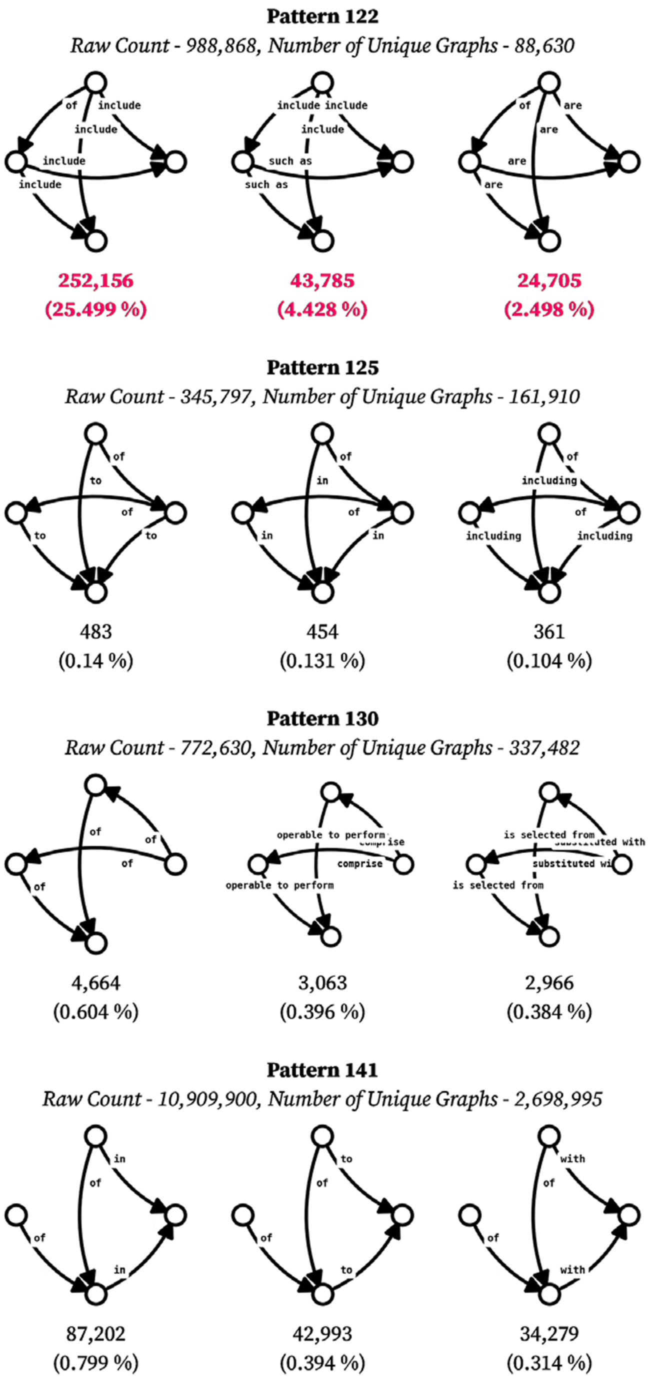

Figure 9. Top three most frequent subgraphs under the motifs represented by Patterns 122, 125, 130 and 141. The frequency of each subgraph and its percentage with respect to raw motif count are mentioned in the (.) beneath.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

Pattern 122 could be more generalizable. Similarly, the most frequent subgraphs under Patterns 13 and 11 – in addition to pattern 122 – offer generalizable interpretations, potentially explaining the foundations of the knowledge of technological artifacts.

As observed in Figure 8, Pattern 13 uses “of” to capture multiple attributes, subsystems or sub-processes of a higher entity in a technological artifact, for example, “temperature **of** the surface.” It also uses “are” to assign specific entities, for example, “the unbound candidate nucleic acids **are** compound A.” Pattern 11 uses “of” to capture attributes of an entity that is specified “in” another, for example, “the pressure **of** the liquid **in** state A.” Pattern 122 is a dominant motif in A61, which involves a majority of the sentences like “examples **of** material **include** compound A, compound B…” that assume hierarchical knowledge structures as in Figure 9. As Pattern 11 is a dominant motif in A61, the domain of chemical compounds for medical applications, we can observe a generalizable subgraph communicating a selection of chemical compounds, for example, “a molecule **selected from** the group **consisting of** alkyls.” Although these relationships are not common (see Figure 5d), their combination as found in Pattern 11 is more common and generalizable. Even though Patterns 13 and 11 appear to represent aggregation [→ · ←] and sequence [→ · →], most generalizable subgraphs, that reveal the actual relationships in the edge labels, tell us that these patterns capture hierarchy [← · →] of entities in artifacts. While Pattern 122 already exhibits a hierarchy, the most frequent subgraphs within these dominant motifs form hierarchical knowledge structures. The motifs and the underlying subgraphs showcase how knowledge is combined at a local level in technological artifact descriptions. The results shown in Figures 8 and 9 indicate that text descriptions fundamentally capture the design hierarchy of technological artifacts.

3.3. Knowledge specification strategies

Our findings upon investigating the linguistic and structural foundations of patent documents suggest that artifact descriptions largely include abstract entities (for example, “the NN,” “JJ NNP”) that are locally connected by abstract relationships (for example, “of,” “in”) as indicated in the subgraphs of dominant motifs (Figures 8 and 9). Such abstract entities, relationships and subgraphs often constitute redundant knowledge structures. In this subsection, we showcase strategies to specify such abstract entities so that we can simplify knowledge structures derived from technological artifact descriptions. As illustrated in Figure 10, “**an array** and **four strings** of **memory cells**” can be captured using relatively complex entities like “memory cell array” and “four memory cell strings,” collapsing the corresponding knowledge structure, that is, Pattern 13. As most frequent relationships (“of,” “in”) lack sufficient context and thus become more mutable (Jamrozik & Gentner 2020), they can be specified, for example, “in” → “observed in” as shown in Pattern 11, Figure 10. Knowledge structures could also be simplified by eliminating redundant “include” edges, as indicated with the “acyl groups” example.

The examples shown in Figure 10 indicate a broad limitation of using natural language in technological artifact descriptions. The example of Pattern 11 in Figure 10 is derived from the text, “**movement m** of **the robot** in **the migration mode**,” which shows that “movement m” and “the robot” are mentioned abstractly

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

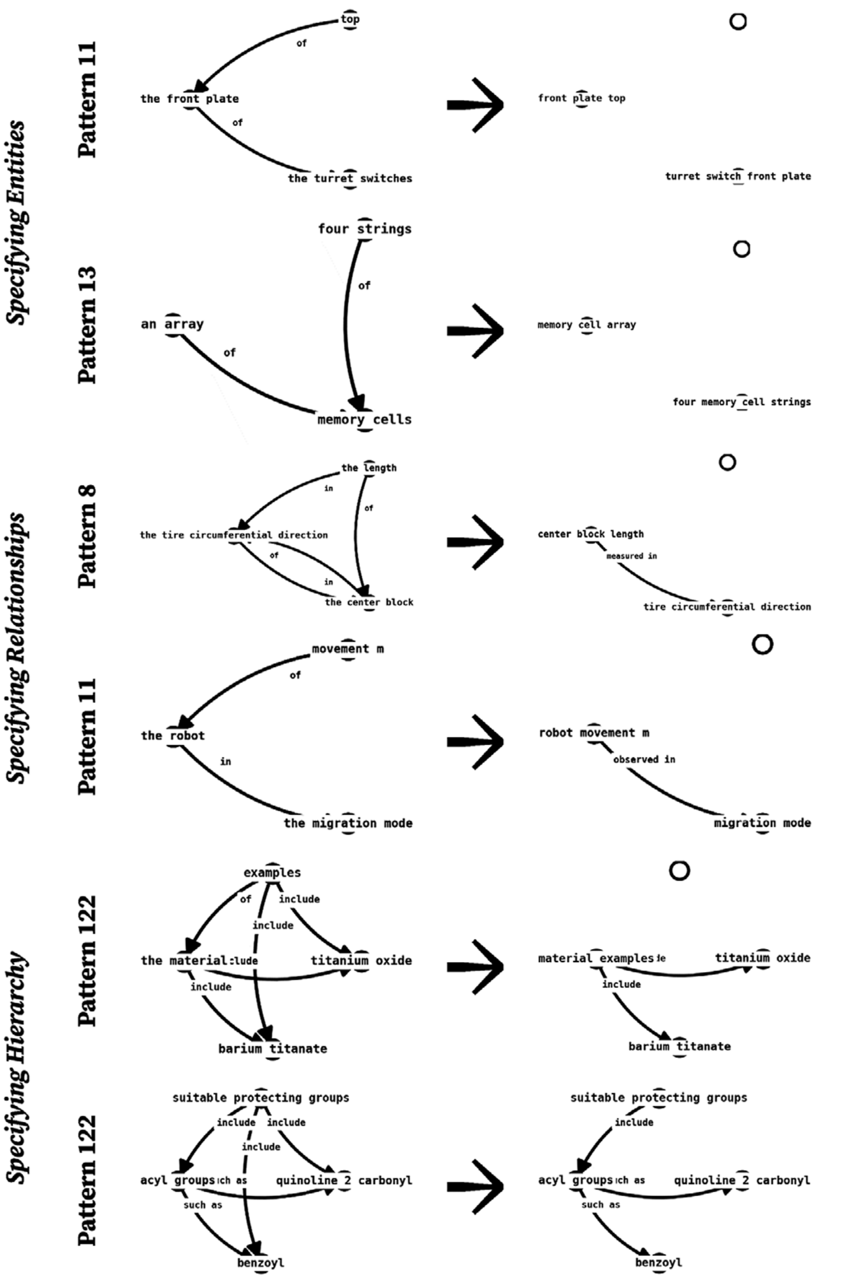

Figure 10. Illustrating specification of knowledge entities (Adjustable tonneau cover - https://patents.google.com/patent/US11084362B2/; Memory configured to perform logic operations on values representative of sensed characteristics of data lines and a threshold data value – https://patents.google.com/patent/US11074982/), relationships (Pneumatic tire for heavy loads – https://patents.google.com/patent/US10308079/; Autonomous mobile robot system – https://patents.google.com/patent/US9229454B1/) and hierarchical structures (Dispersion liquid, composition, film, manufacturing method of film and dispersant – https://patents.google.com/patent/US10928726/; Polar-substituted hydrocarbons – https://patents.google.com/patent/US6071895/).

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

to associate them with other entities of the overall patent artifact – for example, "**the path** and **duration** of **movement m** are better than **movement n** of **the robot**." In this sentence, knowledge is not captured individually through entities such as "path," "duration," "movement" and "robot," but through their verbal and syntactical associations in the sentence. In order to densely pack information within a sentence, it is necessary to keep the entities as simple as possible while introducing several connections among them. As a consequence, we observe that the majority of the entities are abstract nouns (Figure 5) that are associated using abstract relationships (Figures 8 and 9).

The sentence can alternatively be written as, "**the robot movement m path and the robot movement m duration** are better than **the robot movement n path** and **the robot movement n duration**." It is, however, impractical to have artifact knowledge documented in this manner. Alternatively, we can re-represent the knowledge extracted from natural language text, however it is written, so that knowledge constituents such as entities and relationships in the knowledge graphs are less abstract and thus more specific. As illustrated further in Figure 11, the specification strategies can also help modularize the structure of patent knowledge graphs. Effectively modularizing systems can incur various benefits, such as change management, intellectual property protection, technical robustness and others (Baldwin & Henkel 2015; Siddharth & Sarkar 2018, 2017). In the patent literature, Siddharth (2025) observes that modularizing knowledge structures enhances the technological impact of patents.

We apply the knowledge specification strategies to the domain of coffee grinder patents, as shown in Figure 12. For this purpose, we retrieved the full text of seven patents having the title "coffee-grinder." We apply our knowledge graph extraction method (Siddharth & Luo 2024; Siddharth 2024a) to the descriptions of these coffee-grinder patents to extract 3,618 facts. Among these, we filter 320 generalizable facts that have a frequency ≥ 2. In Figure 12-a, we combine the facts around the core entity – "coffee grinder" and visualize them as a knowledge graph. In this knowledge graph, the unlabeled edges indicate a hierarchical relationship – most commonly represented by "include" and alternatively using other forms as indicated in Figure 6. The labelled edges mention a specific relationship between a pair of entities. Hence, both hierarchical and other relationships represent associations among common entities surrounding a coffee grinder. This knowledge graph serves as an explicit representation of generalizable domain knowledge of coffee grinders, as derived from the corresponding patent descriptions.

As we apply the specification strategies, we modify the fact "**receivers** of **frustonical burr**" to "frustonical burr receivers," reducing the level of abstraction in either of the entities "receivers" and "frustonical burr." Similarly, we specify the other entities that are associated with the "of" relationship. In addition, we specify the abstract relationships such as "**coffee grinder** on **firm, horizontal surface**" to "**coffee grinder** rested on **firm, horizontal surface**" based on the context offered by the pair of entities. Upon performing these modifications, as shown in Figure 12b, we find that the structure of the knowledge graph is more simplified and the individual entities/relationships are more enriched in terms of information content. In addition to this case study, we apply these knowledge specification strategies to a relatively more complex glue gun knowledge graph as shown in Appendix C.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

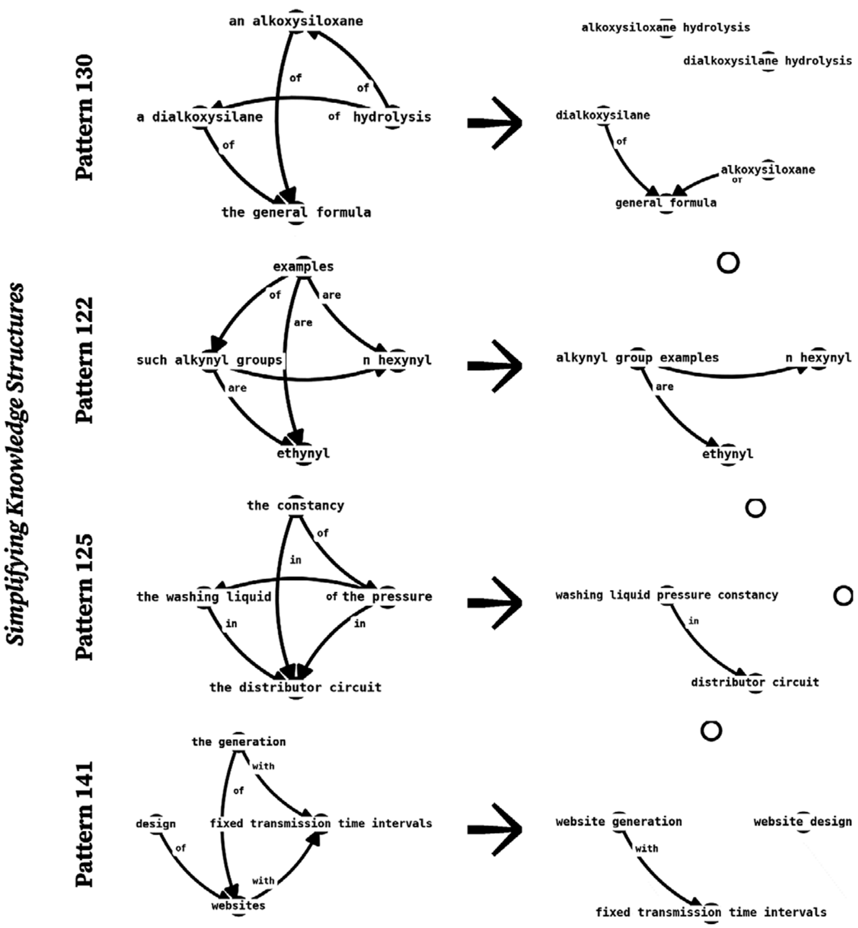

Figure 11. Illustrating simplification of knowledge structures using examples of Pattern 130 (Process for preparing low molecular weight organosiloxane terminated with silanol group – https://patents.google.com/patent/US5576408/), 122 (Polar-substituted hydrocarbons – https://patents.google.com/patent/US6071895/), 125 (Machine for washing objects and method for the hydraulic and mechanical connection of a trolley carrying objects to be washed to a feed circuit of a washing liquid for a machine for washing objects – https://patents.google.com/patent/US9393600/) and 141 (Systems and methods for loading websites with multiple items – https://patents.google.com/patent/US11055378/).

4. Conclusions

In this study, we addressed the question "*What is the basis of the knowledge of technological artifacts?*" from a perspective different from the literature, where similar questions are often addressed through expert-defined ontologies. Drawing on 33,881 utility patent descriptions represented as knowledge graphs, we derived the foundations of artifact knowledge as generalizable linguistic syntaxes and structural motifs. The linguistic syntaxes primarily comprise noun-based entities, abstract relational terms such as "of," "in," and hierarchical variants of "include." These

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

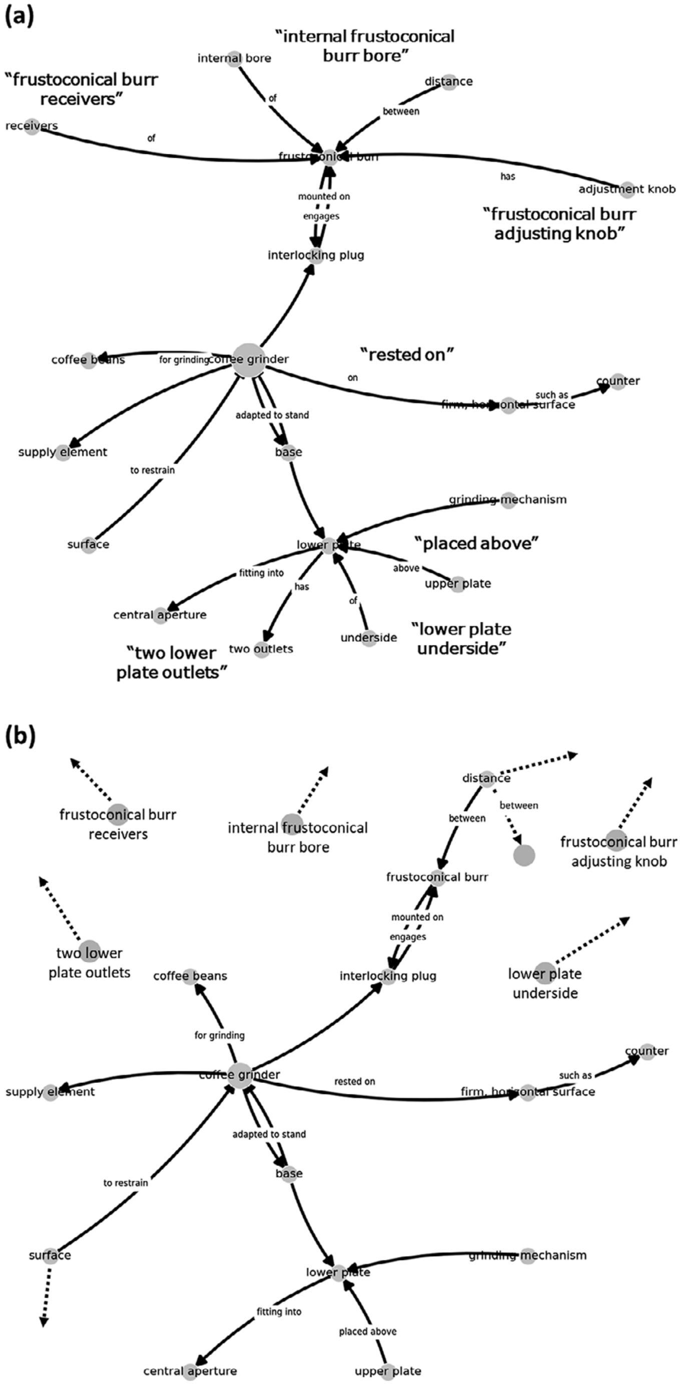

Figure 12. (a) Extended neighborhood of "coffee grinder" entity. The edges without a label indicate a hierarchical relationship – "include." (b) Transformed extended neighborhood of "coffee grinder" entity. The edges without a label indicate a hierarchical relationship – "include."

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

linguistic syntaxes – derived from a large, stratified sample of patents – could be considered the primary linguistic foundations for all terms in the technology space.

Through the motif discovery approaches inspired by studies in biology and ecology, we identified eight unique subgraph patterns that represent how terms are locally combined in technological artifact descriptions. Among these, we converged on Patterns 11, 13 and 122, whose subgraphs offered generalizable interpretations based on their relatively higher frequency proportions. Examining the subgraphs within the selected patterns revealed that artifact descriptions primarily capture the design hierarchy of artifacts, largely through abstract relationships such as "of," "in," "include" and others. These findings could echo throughout the technology space, as we have derived them from a large, stratified sample of patents from the total technology space.

The major conclusions of this study are as follows:

- The knowledge of technological artifacts has simple and abstract linguistic foundations, represented by basic forms of nouns, prepositions ("of," "in") and verbs ("include"), which themselves combine to express complex terms and phrases.
- The knowledge of technological artifacts has non-random and statistically significant structural foundations, represented by a finite number of 3-node and 4-node subgraph patterns, which predominantly capture the design hierarchy of artifacts.
- The knowledge of technological artifacts from patent or similar natural language descriptions largely comprises abstract knowledge at the local level, where (automated) specification strategies are required for knowledge retrieval applications.

The methods and findings in this study make contributions to design literature and can enable applications in design research and practice, such as knowledge engineering, ontology extraction, function modelling, graph-based reasoning, patent search, text-mining, technology mapping and LLM-centric applications. The emphasis of findings on design hierarchy and the level of abstraction can provide an informed perspective for future design studies that leverage natural language artifact descriptions. Knowledge search is quite expensive in design and innovation processes, because of the false positives incurred by the abstract terms in natural language text sources (Han, Zhang, & Tong 2024). Re-representing the knowledge of technological artifacts – supposedly using our specification strategies – could thus circumvent the common issues of knowledge retrieval, enhance the performance and enable trustworthy adoption of LLMs that are being increasingly utilized for knowledge retrieval (Siddharth & Luo 2025).

The limitations of this study pave the way for future research directions. Currently restricted to patents from the USPTO, the study shall explore other important sources of artifact knowledge, such as scientific articles, design reports and more. The findings based on Zipf distribution analyses and motif discovery require tests of robustness, which typically involve replication of findings when patents are studied across domains, time frames, geographies and other variants. Our study must move from sample to population-level analyses to carry out such robustness checks. The proposed knowledge specification strategies are manually demonstrated with a few patents on coffee grinders and glue guns. A comprehensive demonstration with a family of products and services needs to be carried out –

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

preferably by automating specification strategies using LLMs – and studies must be conducted to assess how they aid in knowledge representation and retrieval in the design and innovation processes.

Notes

[1] Glue gun – https://patents.google.com/patent/US4535916A/

[2] Dataset of patent sentences and engineering design facts – https://huggingface.co/datasets/siddharthl1293/engineering_design_facts

[3] US Patent Database access facilitated by Patents View - https://patentsview.org/download/data-download-tables

[4] https://www.qualtrics.com/articles/strategy-research/calculating-sample-size/

[5] Cooperative Patent Classification Scheme – https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html

[6] The steps involved in processing patent text are detailed in Siddharth and Luo (2023, pp. 43, 44).

[7] Examining Patented Artefact Knowledge Graphs to understand Linguistic and Structural Basis. Zenodo. https://doi.org/10.5281/zenodo.13328258

[8] We convert each word in a term into its parts of speech except for the word that includes pronouns, prepositions and 30 most frequent words like “device,” “system,” “said,” “portion,” “information,” “signal,” “invention,” “layer,” “method,” “user,” “material” and others.

[9] SciPy Curve Fit – https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html

[10] The large knowledge graph in our data is extracted from the patent titled, “Method for treating an individual suffering from a chronic infectious disease and cancer” – https://patents.google.com/patent/US11213552B2/

[11] Alon (2006, p. 82) identify 199 patterns in 4-node subgraphs, while our approach based on canonical form identifies only 134 patterns, not distinguishing between a few isomorphs. Our canonical form includes in-degrees, out-degrees and number of edges that are not sufficient to distinguish between 4-node isomorphs that have the same features but different orientations of edges, especially when the edge count is high.

[12] In sensory transcription networks, a node can self-relate as termed, “auto-regulation” (Alon 2006, p. 37), not applicable here.

[13] The raw frequencies of entities and relationships are available externally in the dataset (Siddharth 2024b).

[14] Sentence Transformers – https://github.com/UKPLab/sentence-transformers

[15] The motifs indicated by pattern IDs against each patent are listed with figures in the accessible dataset (Siddharth 2024b).

[16] To calculate this, we ranked the motifs in terms of the number of patents in which they are found statistically significant, that is, based on Equations 5 and 6. Assuming the number of patents follows a normal distribution, we compute the z-score for these and filter the ones having a z-score higher than 2.326, which corresponds to a p-value of 0.01 and 99% confidence level.

[17] We sort motifs in terms of the number of patents both by considering the overall of 33,881 patents as well as within most populated domains. For example, the class GO6F has 3,452 patents. To select motifs in this class, we calculate z-scores by the number of patents only within this class. By making these separate calculations, we can examine if there are motifs in these classes distinguishable from those motifs that are found dominant in the overall sample.

[18] This trait is essential in distributions such as Zipf, where the distribution plot resembles an L-shape, meaning that a few items in the distribution have relatively much higher frequency proportions than the remaining ones. Such items can then be selected as representatives of the distribution. We followed a similar approach in our study of linguistic syntaxes as well.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

References

Ahmed, S. K. 2024 How to choose a sampling technique and determine sample size for research: A simplified guide for researchers. *Oral Oncology Reports* **12**, 100662; doi: 10.1016/j.oor.2024.100662.

Alon, U. 2006 *An Introduction to Systems Biology: Design Principles of Biological Circuits*, 1st Edn. Chapman and Hall/CRC. https://doi.org/10.1201/9781420011432.

Baldwin, C. Y. & **Henkel, J.** 2015 Modularity and intellectual property protection. *Strategic Management Journal* **36** (11), 1637–1655; doi:10.1002/smj.2303.

Brennecke, J., **Sofka, W.**, **Wang, P.** & **Rank, O. N.** 2021 How the organizational design of R&D units affects individual search intensity – A network study. *Research Policy* **50** (5), 104219; doi:10.1016/j.respol.2021.104219.

Browning, T. R. 2001 Applying the design structure matrix to system decomposition and integration problems: a review and new directions. *IEEE Transactions on Engineering Management* **48** (3), 292–306; doi:10.1109/17.946528.

Chandrasegaran, S. K., **Ramani, K.**, **Sriram, R. D.**, **Horváth, I.**, **Bernard, A.**, **Harik, R. F.** & **Gao, W.** 2013 The evolution, challenges, and future of knowledge representation in product design systems. *Computer-Aided Design* **45** (2), 204–228; doi:10.1016/j.cad.2012.08.006.

Chandrasekaran, B. & **Josephson, J. R.** 2000 Function in Device Representation. *Engineering with Computers* **16** (3-4), 162–177; doi:10.1007/s003660070003.

Chomsky, N. 1965. Methodological preliminaries, In *Aspects of the Theory of Syntax*. The MIT Press, pp. 1–66. http://www.jstor.org/stable/j.ctt17kk81z.5

De Kleer, J. & **Brown, J. S.** 1984 A qualitative physics based on confluences. *Artificial Intelligence* **24** (1-3), 7–83; doi:10.1016/0004-3702(84)90037-7.

Fantoni, G., **Coli, E.**, **Chiarello, F.**, **Apreda, R.**, **Dell'Orletta, F.**, & **Pratelli, G.**, 2021. Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector. *Computers in Industry* **124**, 103357; doi:10.1016/j.compind.2020.103357

Fenves, S. J., **Foufou, S.**, **Bock, C.** & **Sriram, R. D.** 2008 CPM2: A core model for product data. *Journal of Computing and Information Science in Engineering* **8** (1); doi:10.1115/1.2830842.

Fu, K., **Cagan, J.**, **Kotovsky, K.** & **Wood, K.** 2013 Discovering structure in design databases through functional and surface based mapping. *Journal of Mechanical Design* **135** (3); doi:10.1115/1.4023484.

Gruber, T. R. 1993 A translation approach to portable ontology specifications. *Knowledge Acquisition* **5** (2), 199–220; doi:10.1006/knac.1993.1008.

Hagoort, P. 2019 The neurobiology of language beyond single-word processing. *Science* **366** (6461), 55–58; doi:10.1126/science.aax0289.

Han, N., **Zhang, Y.** & **Tong, T. W.** 2024 Appropriability risk and knowledge search on digital platforms. *Research Policy* **53** (7), 105028; doi:10.1016/j.respol.2024.105028.

Hirtz, J., **Stone, R. B.**, **McAdams, D. A.**, **Szykman, S.** & **Wood, K. L.** 2002 A functional basis for engineering design: Reconciling and evolving previous efforts. *Research in Engineering Design* **13** (2), 65–82; doi:10.1007/s00163-001-0008-3.

Jamrozik, A. & **Gentner, D.** 2020 Relational labeling unlocks inert knowledge. *Cognition* **196**, 104146; doi:10.1016/j.cognition.2019.104146.

Li, G., **Siddharth, L.** & **Luo, J.** 2023 Embedding knowledge graph of patent metadata to measure knowledge proximity. *Journal of the Association for Information Science and Technology* **74** (4), 476–490; doi:10.1002/asi.24736.

Liu, W., **Tao, Y.** & **Bi, K.** 2022 Capturing information on global knowledge flows from patent transfers: An empirical study using USPTO patents. *Research Policy* **51** (5), 104509; doi:10.1016/j.respol.2022.104509.

Malhotra, A., **Zhang, H.**, **Beuse, M.** & **Schmidt, T.** 2021 How do new use environments influence a technology's knowledge trajectory? A patent citation network analysis of lithium-ion battery technology. *Research Policy* **50** (9), 104318; doi:10.1016/j.respol.2021.104318.

Milo, R., **Shen-Orr, S.**, **Itzkovitz, S.**, **Kashtan, N.**, **Chklovskii, D.** & **Alon, U.** 2002 Network Motifs: simple building blocks of complex networks. *Science* **298** (5594), 824–827; doi:10.1126/science.298.5594.824.

Mollica, F. & **Piantadosi, S. T.** 2019 Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science* **6** (3), 181393; doi:10.1098/rsos.181393.

Murphy, J., **Fu, K.**, **Otto, K.**, **Yang, M.**, **Jensen, D.** & **Wood, K.** 2014 Function based design-by-analogy: a functional vector approach to analogical search. *Journal of Mechanical Design* **136** (10); doi:10.1115/1.4028093.

Ribeiro, P., **Paredes, P.**, **Silva, M. E. P.**, **Aparicio, D.** & **Silva, F.** 2022 A survey on subgraph counting. *ACM Computing Surveys* **54** (2), 1–36; doi:10.1145/3433652.

Robinson, M., **Taube-Adams, B.**, **Kang, S.** & **Dong, A.** 2024 A functional perspective on the emergence of dominant designs. *Journal of Mechanical Design* **146** (5); doi:10.1115/1.4064043.

Rodenacker, W. G. & **Rodenacker, W. G.** 1976 Methodisches Konstruieren. In *Methodisches Konstruieren* (ed. W. G. Rodenacker), pp. 36–224. Springer. 10.1007/978-3-662-08721-3_2.

Siddharth, L. 2024a Training infrastructure for engineering design knowledge extraction; Accessed date: June, 2025. doi:10.5281/zenodo.12012131

Siddharth, L. 2024b Examining patented artefact knowledge graphs to understand linguistic and structural basis. Accessed date: June, 2025. 10.5281/zenodo.13328258

Siddharth, L. 2025 Modularizing patent knowledge for enhanced technological impact. *World Patent Information* **83**, 102407; doi:10.1016/j.wpi.2025.102407.

Siddharth, L. & **Luo, J.** 2023 Towards populating generalizable engineering design knowledge. *arXiv (cs.CL).* Accessed date: June, 2025. Available at: https://arxiv.org/abs/2307.06985v3

Siddharth, L. & **Luo, J.** 2024 Retrieval augmented generation using engineering design knowledge. *Knowledge-Based Systems* **303**, 112410; doi:10.1016/j.knosys.2024.112410.

Siddharth, L. & **Luo, J.** 2025 Data-driven innovation for trustworthy AI. *She Ji: The Journal of Design, Economics, and Innovation* **11** (3), 261–283; doi:10.1016/j.sheji.2025.06.002.

Siddharth, L. & **Sarkar, P.** 2017 A methodology for predicting the effect of engineering design changes. *Procedia CIRP* **60**, 452–457; doi:10.1016/j.procir.2017.03.071.

Siddharth, L. & **Sarkar, P.** 2018 A multiple-domain matrix support to capture rationale for engineering design changes. *Journal of Computing and Information Science in Engineering* **18** (2); doi:10.1115/1.4039850.

Siddharth, L., **Chakrabarti, A.** & **Ranganath, R.** 2020 Modeling and structuring design rationale to enable knowledge reuse. *Systems Engineering* **23** (3), 294–311; doi:10.1002/sys.21515.

Siddharth, L., **Blessing, L. T. M.**, **Wood, K. L.** & **Luo, J.** 2022 Engineering knowledge graph from patent database. *Journal of Computing and Information Science in Engineering* **22** (2); doi:10.1115/1.4052293.

Siddharth, L., **Blessing, L.** & **Luo, J.** 2022a Natural language processing in-and-for design research. *Design Science* **8**, e21; doi:10.1017/dsj.2022.16.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

Siddharth, L., **Li, G.** & **Luo, J.** 2022b Enhancing patent retrieval using text and knowledge graph embeddings: a technical note. *Journal of Engineering Design* **33** (8-9), 670–683; doi:10.1080/09544828.2022.2144714.

Simon, H. A. 1962 The architecture of complexity. *Proceedings of the American Philosophical Society* **106**, 467–482; http://www.jstor.org/stable/985254.

Stone, L., **Simberloff, D.** & **Artzy-Randrup, Y.** 2019 Network motifs and their origins. *PLoS Computational Biology* **15** (4), e1006749; doi:10.1371/journal.pcbi.1006749.

Strona, G., **Nappo, D.**, **Boccacci, F.**, **Fattorini, S.** & **San-Miguel-Ayanz, J.** 2014 A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications* **5** (1), 4114; doi:10.1038/ncomms5114.

Watts, D. J. & **Strogatz, S. H.** 1998 Collective dynamics of 'small-world' networks. *Nature* **393** (6684), 440–442; doi:10.1038/30918.

Zhang, Z., **Zhang, J.** & **Wang, P.** 2024 Measurement of disruptive innovation and its validity based on improved disruption index. *Scientometrics* **129** (11), 6477–6531; doi: 10.1007/s11192-024-05134-9.

Appendix A

Table A1. List of parts-of-speech tags

Tag	Name	Definition	Example
JJ	Adjective	Describes a noun or pronoun	A red car passed
JJR	Adjective	comparative	Shows comparison between two things
JJS	Adjective	superlative	Shows highest degree of quality
RB	Adverb	Modifies a verb, adjective or adverb	She runs quickly
RBR	Adverb	comparative	Compares actions or qualities
RBS	Adverb	superlative	Shows extreme degree of manners
CD	Cardinal number	Expresses quantity or number	He has three books
CC	Coordinating conjunction	Joins words or phrases of equal status	She ran and jumped
DT	Determiner	Introduces a noun and limits its meaning	The dog barked
EX	Existential there	Introduces the existence of something	There is a problem
FW	Foreign word	A word borrowed from another language	The word bona fide applies
UH	Interjection	Expresses emotion or reaction	Oh that hurts

Continued

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

Table A1. Continued

Tag	Name	Definition	Example
LS	List item marker	Marks items in a list	A first B second
MD	Modal	Expresses necessity, possibility or ability	She can swim
NN	Noun	singular or mass	Names a person, place or thing
NNS	Noun	plural	Plural form of a noun
RP	Particle	Forms part of a phrasal verb	Pick up the box
PRP	Personal pronoun	Refers to a person or thing	I saw her
POS	Possessive ending	Shows ownership or possession	John's book
PRP$	Possessive pronoun	Shows ownership for a pronoun	This is my bag
WP$	Possessive wh-pronoun	Wh pronoun showing possession	Whose bag fell
PDT	Predeterminer	Appears before a determiner	All the students came
IN	Preposition or subordinating conjunction	Shows relation or introduces a clause	She sat on chair
NNP	Proper noun	singular	Specific name of a person or place
NNPS	Proper noun	plural	Plural form of proper nouns
SYM	Symbol	Represents a symbol or formula	Value equals x
TO	to	Marks infinitive verbs or direction	She wants to go
VB	Verb	base form	Base form of a verb
VBD	Verb	past tense	Verb in the past tense
VBG	Verb	gerund or present participle	Verb ending in ing
VBN	Verb	past participle	Past participle form of a verb
VBP	Verb	non-third person singular present	Present tense not third person
VBZ	Verb	third person singular present	Present tense for third person
WRB	Wh-adverb	Adverb used in questions	Where did he go
WDT	Wh-determiner	Introduces a noun with a question	Which book helps
WP	Wh-pronoun	Pronoun used in questions	Who arrived early

The tags and names are taken from the Penn Tree Bank POS tag list (POS Tags – https://www.ling.upenn.edu/courses/Fall_2003/ling001/penntreebank_pos.html).

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

Appendix B

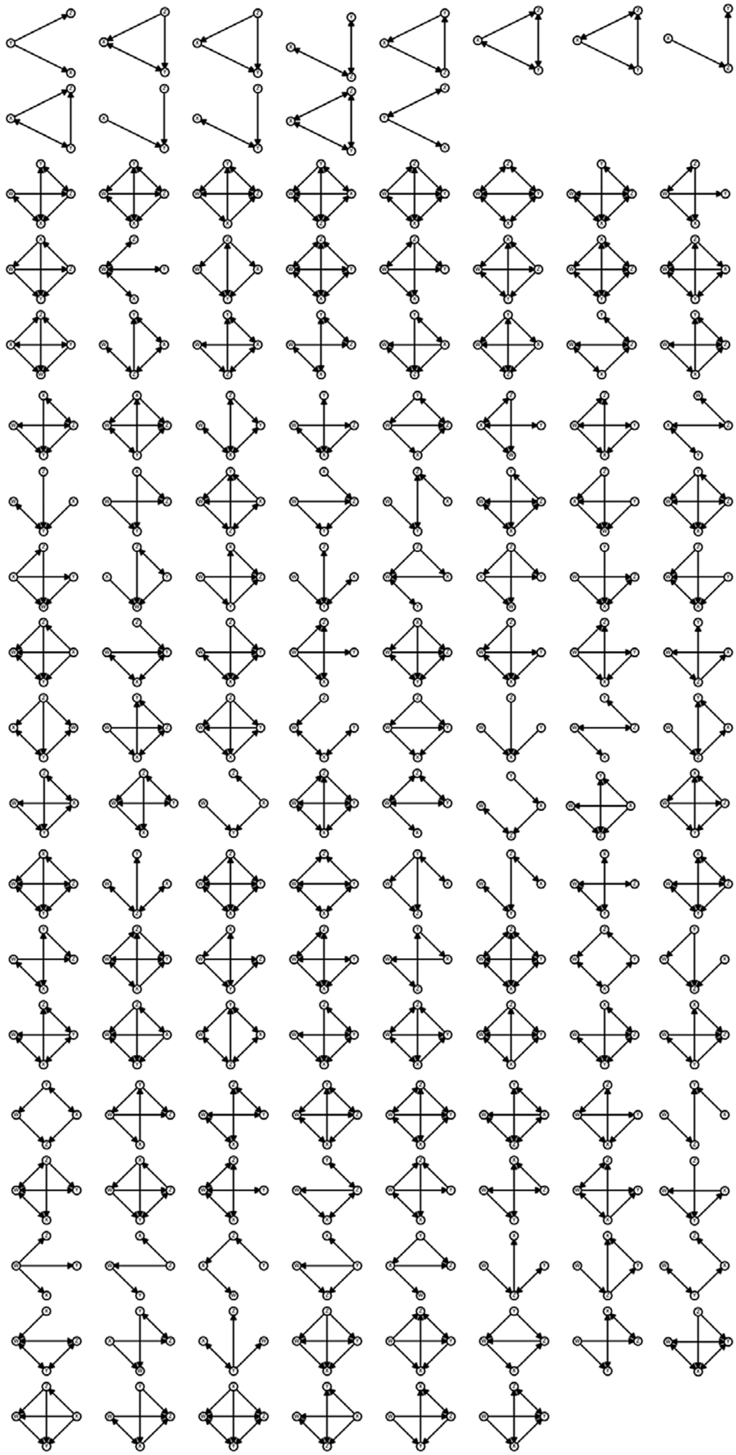

Figure B1. All 3-node and 4-node subgraph patterns.

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press

Appendix C

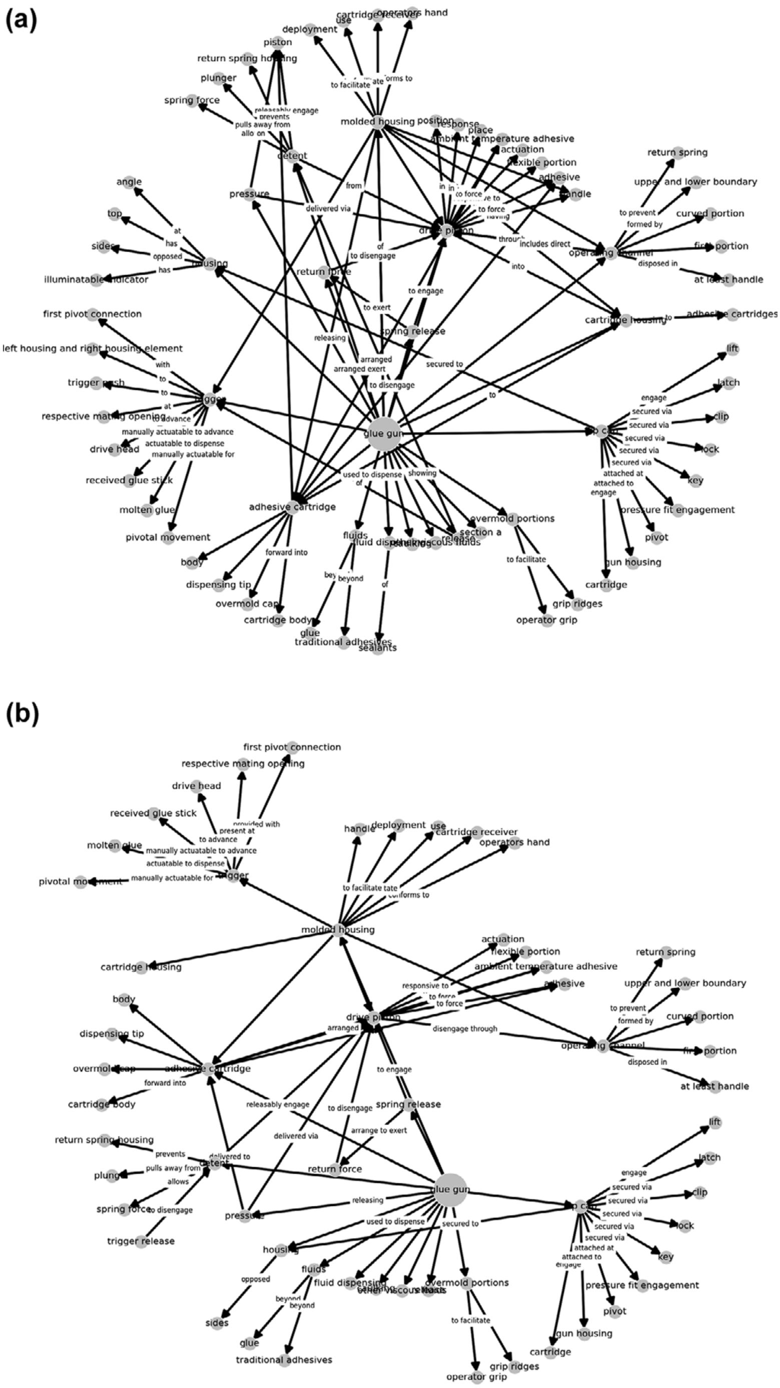

Figure C1. (a) Extended neighborhood of "glue gun" entity. The edges without a label indicate a hierarchical relationship – "include." (b) Transformed extended neighborhood of glue gun. The edges without a label indicate a hierarchical relationship – "include."

https://doi.org/10.1017/dsj.2026.10064 Published online by Cambridge University Press