What can a Multidimensional Language Model Tell Us about Architecture?

Julia Krasselt

Why look at language in architecture?

Language and architecture have much in common. Both function as semiotic systems—they convey meaning beyond their immediate form.[1] Just as a building can be "read" as a text that structures and organizes space,[2] language serves as a fundamental medium for communication. Both language and architecture are inherently social practices: They shape and are shaped by human interaction, structuring communication, behavior, and collective meaning within cultural and societal contexts. Moreover, language and architecture interact with each other. Research in linguistics has shown that the way we conceptualize and communicate about space is shaped by language, with different languages providing different spatial reference frameworks.[3] When people interact with architectural spaces, their perceptions are shaped by linguistic categories and culturally-specific spatial lexicons. Conversely, space and architecture shape the way we speak and communicate. Even in everyday interac-

1 Umberto Eco, *A Theory of Semiotics* (Indiana University Press, 1976); Gabriele Aroni, "Semiotics in Architecture and Spatial Design," in *Bloomsbury Semiotics Volume 2: Semiotics in the Natural and Technical Sciences*, ed. Jamin Pelkey and Stéphanie Walsh Matthews (Bloomsbury, 2023), 277–96, https://doi.org/10.5040/9781350139350.

2 Bryan Lawson, *Language of Space* (Taylor and Francis, 2007), 6.

3 Barbara Landau and Ray Jackendoff, "What and Where in Spatial Language and Spatial Cognition," *Behavioral and Brain Sciences 16*, no. 2 (June 1993): 255–65, https://doi.org/10.1017/S0140525X00029927; Peggy Li and Lila Gleitman, "Turning the Tables: Language and Spatial Reasoning," *Cognition 83*, no. 3 (April 2002): 265–94, https://doi.org/10.1016/S0010-0277(02)00009-4.

tions, space becomes a communicative element: "When we talk to each other, the space between us is part of our communication."[4]

But the relationship between language and architecture is also a complex one. As a physical object, architecture poses a challenge to language because it is primarily experienced through sight, touch, hearing, and smell. The material properties of a building, surface textures, spatial acoustics, and atmospheric qualities all contribute to architectural perception. These elements influence human interaction with the built environment but are not easily translated into verbal descriptions. In addition, architecture involves movement and embodied interaction, adding complexity to linguistic representation.

Yet, language plays a crucial role in conceptualizing, communicating, and theorizing about architecture and architectural practice. Without language, it would be difficult to communicate design intent or analyze the social and cultural roles of architecture. Architectural theorist Tom Markus emphasizes this by stating that "language is at the core of building, using and understanding buildings."[5] Language also structures architectural discourse, allowing professionals to communicate ideas systematically. As the architectural theorist Branko Mitrovic notes, certain aspects of architecture, such as function, social history, or cultural role, are primarily verbal and cannot be visualized.[6]

This article explores the relationship between language and architecture by using word embeddings, a computational method for representing word meanings based on their contextual relationships.[7] Word Embeddings are a fundamental component of Large Language Models (LLMs), serving as the representational backbone that allows these models to efficiently process, understand, and generate human language. Through an exploratory analysis, this study demonstrates how word embeddings can be used to investigate the meaning of key architectural concepts at a given point in time.

The research presented here follows a discourse-analytic perspective. Building on constructivist theories, this study considers language not as

4 Lawson, *Language of Space*, 8.

5 Thomas A. Markus, *Buildings and Power: Freedom and Control in the Origin of Modern Building Types* (Routledge, 1993), 4.

6 Branko Mitrovic, "Architectural Formalism and the Demise of the Linguistic Turn," *Log* 17 (2009): 20.

7 Tomas Mikolov et al., "Distributed Representations of Words and Phrases and Their Compositionality," *Advances in Neural Information Processing Systems 26* (2013): 3111–19; Alessandro Lenci, "Distributional Models of Word Meaning," *Annual Review of Linguistics* 4, no. 1 (2018): 151–71, https://doi.org/10.1146/annurev-linguistics-030514-125254.

a neutral medium reflecting an objective reality but as an instrument that actively constructs it.[8] The way society discusses architecture shapes collective knowledge and influences architectural perception. From a linguistic perspective, discourse analysis involves examining the meaning of words—semantics—to understand how concepts are constructed. By exploiting the power of word embeddings to map semantic relationships, this study offers a novel approach to analyzing architectural discourse.

Word Embeddings

Word embeddings operationalize the distributional hypothesis, a linguistic theory which proposes that the meaning of a word is derived from the contexts in which it occurs.[9] This approach to semantics was established long before the advent of LLMs and Generative AI, both of which rely heavily on it. Instead, distributional approaches to semantics were already being expressed in the 1950s, as the following quotes from Wittgenstein, Firth, and Harris show:

> [T]he meaning of a word is its use in the language.[10]
> You shall know a word by the company it keeps.[11]
> … difference of meaning correlates with difference in distribution.[12]

The distributional hypothesis allows researchers to model the semantic similarity of words, phrases, and even larger syntactic structures based on real-world language usage, rather than relying solely on (native) speakers' intuition. According to the distributional hypothesis, meaning emerges from patterns of

8 Jürgen Spitzmüller and Ingo H. Warnke, "Discourse as a 'Linguistic Object': Methodical and Methodological Delimitations," *Critical Discourse Studies 8*, no. 2 (May 2011): 75–94, https://doi.org/10.1080/17405904.2011.558680; James Paul Gee, An Introduction to Discourse Analysis: Theory and Method, 4th ed. (Routledge, 2014).

9 Magnus Sahlgren, "The Distributional Hypothesis," Italian Journal of Linguistics 20 (2008): 33–53.

10 Ludwig Wittgenstein, *Philosophical Investigations*, trans. G.E.M. Anscombe (Basil Blackwell, 1958), PU §43.

11 John Rupert Firth, *Papers in Linguistics 1934–1951* (Oxford University Press, 1957).

12 Zellig S. Harris, "Distributional Structure," WORD 10, no. 2–3 (August 1954): 156, https://doi.org/10.1080/00437956.1954.11659520.

word co-occurrence. Thus, words that frequently occur in similar linguistic environments tend to have related meanings. Rather than considering contextual features such as time or place, this approach focuses on co-text—the words that commonly surround a given term.

Contextual patterns can be systematically identified in large collections of text. While early theories conceptualized meaning through linguistic context, modern AI techniques have operationalized this principle computationally.[13] Word embeddings model semantic similarity by mapping words into high-dimensional vector spaces: each word from a given corpus is represented as a vector in such a space. The position of each word in this space is determined by the surrounding words with which it appears. The following sentence from a corpus of Swiss architectural magazines (see the following section) shows the word *Beton* ("concrete") in its immediate context:

left context (10 words)	node	right context (10 words)	source text id
Stahlbeton eingebaut und die hangseitige Wand mit einem Vorbau aus	*Beton*	*versehen. Der neue Betonkubus unterteilt den Keller in eine*	Schweizer Bauzeitung, 142/2016
"Reinforced concrete was installed and the wall on the slope side was covered with"	"concrete"	"The new concrete cube divides the basement into a"	

A word such as *Beton* is likely to be repeatedly embedded in similar contexts, meaning that one would find similar sentences like those shown above, giving rise to specific patterns of use. Large linguistic corpora are particularly effective in identifying these patterns. They not only allow the identification of patterns for individual words but also reveal patterns that emerge from clusters of words. Put differently, by applying such a distributional, context-based approach to all the words in a corpus, they can be clustered according to their similar contexts. This is achieved by constructing a co-occurrence matrix that captures word-context relationships, as shown in the following simplified example:

13 Lenci, "Distributional Models of Word Meaning."

	Gebäude ("building")	*bauen* ("to build")	*malen* ("to paint")	*skulptural* ("sculptural")	*Wendel-treppe* ("spiral staircase")	…
Beton ("concrete")	6	5	1	1	1	…
Holz ("wood")	5	7	3	0	0	…
Gemälde ("painting")	2	1	7	0	0	…
Porträt ("portrait")	1	2	8	0	0	…

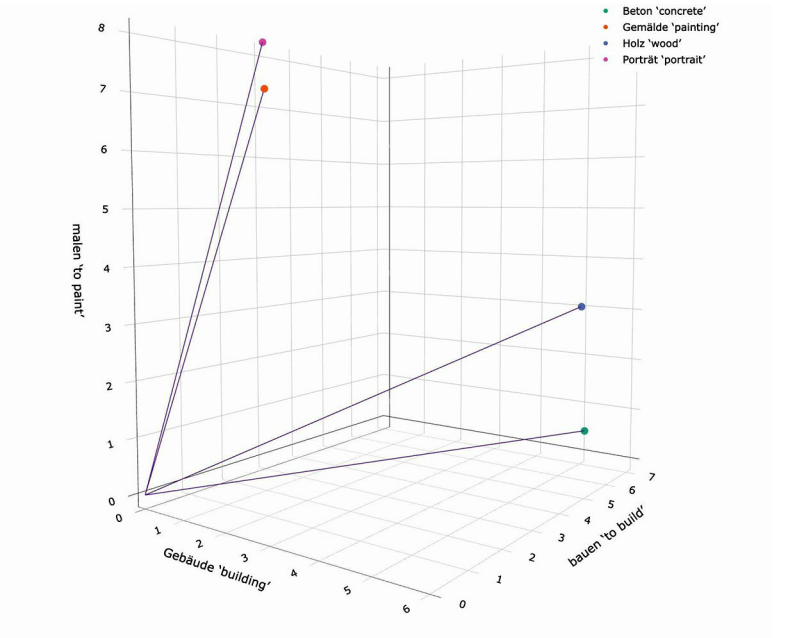
By representing words as numerical vectors, word embeddings allow semantic similarity to be measured mathematically. For example, words that occur in similar contexts (e.g., *Beton* and *Holz*) will have similar vectors and will be positioned closer together in the vector space (see fig. 20). In practical implementations, word embedding models go beyond simple co-occurrence matrices. Using neural network-based learning techniques, models such as Word2Vec, GloVe, and FastText generate vector representations that capture complex semantic relationships.[14] These models compute word embeddings with hundreds of dimensions, allowing them to reflect nuanced word meanings. While co-occurrence matrices provide valuable insights, they are limited by sparsity and dimensional constraints. Neural network-based models overcome these limitations by producing dense vector representations that capture more nuanced semantic relationships.

A simplified 3D representation can be used to visualize word embeddings. In such a representation, words with similar contexts will be closer together, while unrelated words will be further apart (typically measured by cosine similarity or Euclidean distance). Fig. 21 illustrates how words such as *Beton* ("concrete"), *Holz* ("wood"), *Gemälde* ("painting"), and *Porträt* ("portrait") are

14 Piotr Bojanowski et al., "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics 5 (June 1, 2017): 135–46, https://doi.org/10.1162/tacl_a_00051; Mikolov et al., "Distributed Representations of Words and Phrases and Their Compositionality"; Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), ed. Alessandro Moschitti, Bo Pang, and Walter Daelemans (Association for Computational Linguistics, 2014), 1532–43, https://doi.org/10.3115/v1/D14-1162.

positioned in a vector space based on their contextual similarity. For example, the vectors for *Beton* and *Holz* are closer together than *Holz* and *Gemälde*, indicating a stronger contextual similarity between building materials. This method allows for accurate modelling of word relationships, even in high-dimensional spaces.

Fig. 19: 3D vector space as visualization for word embeddings.



Nearest neighbor analysis in a word embedding model is a powerful tool for semantic analysis, revealing how meaning is structured within a corpus. It allows researchers to empirically identify dominant semantic fields without relying on manually-assigned categories, providing a data-driven linguistic perspective. In addition, nearest neighbor analysis helps uncover conceptual metaphors and highlight patterns in language use that reflect underlying

discourse structures.[15] In the following analysis, this method is applied to examine how the architectural concept of spatiality is used within specific discursive contexts, namely within architectural magazines.

Data

For the study, a corpus of three Swiss architectural journals was compiled: *Werk, Bauen und Wohnen* (1977–2021), *Schweizerische Bauzeitung* (now *Tec 21*) (2001–2017), and *Hochparterre* (1988–2022). All issues of these journals were provided by the ETH Zurich Library in a digitized XML format. The corpus includes all textual elements from the three journals (e.g., articles and image captions) but excludes images (see Table 1).

Table 1: Corpus of Swiss architectural journals.

	words	documents	time span
Werk, Bauen und Wohnen	17.96 million	10,418	1977–2021
Schweizerische Bauzeitung/Tec21	6.76 million	4,231	2001–2017
Hochparterre	20.4 million	10,092	1988–2022
Total	45.12 million	24,741	

The corpus was processed using an automated linguistic processing pipeline[16] and contains various linguistic annotations (e.g., word and sentence boundaries, parts of speech) as well as text-based metadata (e.g., publication date and issue number). For the corpus, a word-embedding model was computed using word2vec (100 dimensions, context window size of 5).

15 Austin C. Kozlowski, Matt Taddy, and James A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," *American Sociological Review* 84, no. 5 (October 2019): 905–49, https://doi.org/10.1177/0003122419877135.

16 Described in Julia Krasselt et al., "Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (European Language Resources Association, 2020), 4145–51.

The Concept of Spatiality in Architectural Magazines

Spatiality is a central concept in architectural discourse, making it an ideal focus for this exploratory analysis. To analyze how spatiality is represented linguistically, the nearest neighbors of the words *Raum* ("space") and *räumlich* ("spatial") were examined in the word embedding model (Fig. 21). By manually clustering the nearest neighbors, we identified different linguistic dimensions of spatiality in architectural discourse.

1) Functional Spaces

Nearest neighbors in this category refer to spaces with a specific use or purpose in architecture, emphasizing their designed function. Examples include *Wohnraum* ("living space"), *Stadtraum* ("urban space"), *Kirchenraum* ("church space"), *Arbeitsraum* ("workspace"), and *Bewegungsraum* ("movement space"). These terms highlight how space is structured and given meaning through its functional role in architectural contexts.

2) Social Spaces and Interaction

This category includes terms that emphasize space as a site of human interaction and social exchange. Examples include *Begegnungsort* ("meeting place"), *Kommunikationszone* ("communication zone"), *Gemeinschaftsbereich* ("community area"), and *Bewegungsraum* ("movement space"). The presence of these terms suggests that spatiality in architecture is not only physical but also relational, shaped by social dynamics and interaction.

3) Private and Intimate Spaces

Certain terms emphasize the personal or secluded aspects of space, emphasizing privacy and intimacy. Examples include *Privatraum* ("private space"), *Intimität* ("intimacy"), *Geborgenheit* ("sense of security"), and *Privatheit* ("privacy"). These words suggest that spatiality can also be framed in terms of emotional and psychological experiences, emphasizing the importance of enclosed, protective environments.

Fig. 20: Screenshot of word-embedding model calculated for the three Swiss architectural journals. The screenshot shows the 100 nearest neighbors for the word räumlich, "spatial."

4) Spatial Boundaries and Containment

Some words define space in terms of enclosure or extension, reflecting different ways of structuring space. Examples are *Innenraum* ("interior space"), *Aussenraum* ("exterior space"), and *Umraum* ("surrounding space"). Within this category, we also found the metaphors *space as a stage* (*Bühne* "stage," *Szenerie* "scenery", *dramaturgisch* "dramaturgical") and *space as a container* (*Gefäß* "vessel," *Leere* "emptiness"). These findings suggest that architectural discourse often conceptualizes space in terms of containment and permeability. Furthermore, the blurring of boundaries (*durchlässig* "permeable," *fliessender Übergang* "smooth transition," *verschränken* "entangle") reflects a linguistic tendency to describe space as fluid rather than rigidly bound.

5) Spatial Relationships and Connectivity

Words in this category describe how spaces are linked, interwoven, or structured in relation to each other. Examples include *Verflechtung* ("interweaving"),

Verknüpfung ("connection"), *Überlagerung* ("superimposition"), *Raumgefüge* ("spatial structure"), *symbiotisch* ("symbiotic"), and *kompositorisch* ("compositional"). This cluster suggests that space in architecture is not static but dynamic, with an emphasis on relationships between spatial elements rather than isolated units.

6) Sensory and Multimodal Aspects of Space

Beyond its geometric and functional properties, spatiality is also described in experiential and sensory terms. Examples include *Atmosphäre* ("atmosphere"), *Raumerlebnis* ("spatial experience"), *Raumgefühl* ("sense of space"), *erlebbar* ("perceivable"), *spannend* ("tense"), and *klanglich* ("acoustic"). These words indicate that spatiality in architectural discourse extends beyond the visual and geometric to include atmospheric and sensory dimensions, in line with phenomenological perspectives on space.

The identified semantic categories reveal that architectural discourse constructs spatiality as a multidimensional concept, encompassing functionality, social interaction, sensory experience, and dynamic relationships. This analysis provides insights into how space is linguistically framed and conceptualized, reflecting broader architectural thinking. A notable observation is the absence of regulatory and political aspects in the nearest neighbors of *Raum* ("space") and *räumlich* ("spatiality"). While spatial planning, zoning laws, and urban policies play a crucial role in shaping architecture, these aspects do not feature prominently in the linguistic model for architectural journals. This suggests that spatiality in the analyzed architectural journals primarily reflects descriptive and experiential meanings rather than regulatory language.

The exploratory analysis focuses on the linguistic representation of spatiality. The findings open up further research avenues in architectural, sociological, and philosophical contexts. For example, future work could explore how these linguistic structures align with architectural theories of space and whether similar patterns emerge in other domains of architectural discourse.

Outlook

The findings of this study demonstrate how a linguistic approach using word embeddings can reveal underlying semantic structures in architectural discourse. By analyzing the nearest neighbors of *Raum* ("space") and *räumlich*

("spatial") in a large corpus of Swiss architectural journals, we identified different linguistic dimensions of spatiality, including functional spaces, social interaction, private spaces, spatial boundaries, spatial relationships, and sensory experience. This analysis reveals how spatiality is constructed in discourse and provides new insights into the way language encodes architectural concepts.

For architectural research, this study provides a framework for understanding how linguistic patterns reflect and shape architectural thought. While architects are primarily concerned with visual and material forms, their discourse is inherently structured by language. The semantic categories identified in this study suggest that certain aspects of space—such as functionality and connectivity—are linguistically dominant, while regulatory and political dimensions are less prominent in the analyzed corpus. This raises questions about how different genres of architectural writing (e.g., policy documents, academic texts, practitioner discourse) construct space differently.

These findings suggest several avenues for future research, such as how architectural discourse changes over time. By applying diachronic word embedding models, researchers could trace shifts in spatial conceptualization over time, revealing how the meanings of spatial terms evolve in response to architectural trends and societal changes. Another way forward could be to integrate linguistic analysis with architectural theory. While this study focuses on linguistic structures, a next step could be to systematically link the identified semantic categories to architectural and spatial theories (e.g., phenomenological perspectives in architecture).