Ordinary Meaning in Legal Interpretation: A Proposal from a Corpus (and a Bit of an LLM) Perspective

by

Stefan Th. Gries*

Received June 20, 2025

This paper first surveys a variety of problems endemic in the notion of ordinary meaning and then argues in favor of a semantic approach to ordinary meaning that is intensionalist, prototype-based, and corpus-linguistic in nature. This approach uses embeddings models rather than large language models (LLMs), which makes it more replicable, versatile (it can be applied to words that did not exist at corpus time), and more cheaply/quickly adoptable than the current traditional practice. I exemplify the approach using two embeddings models – one from a 2014 Common Crawl of the WWW, the other trained on 1950s American English corpus data.

Keywords: ordinary meaning, prototypes, corpus linguistics, intensional semantics, embeddings

JEL classification code: K1, Z0

1 Ordinary Meaning and/in Legal Interpretation

1.1 Ordinary Meaning, Dictionaries, and Extensional Corpus Approaches

From an expository or didactically/rhetorically useful perspective, we can distinguish two different positions regarding complexity and intuition in legal decision making. On the one hand, lawyers respond to the increasing complexity of the decision problem with a reduction, not an increase, in the complexity of the process, which could be simple (doctrinal) rules for a complex world. On the other hand, one could argue that the human mind is actually surprisingly well prepared to handle complexity and even indeterminacy but that it does so below the surface of consciousness with intuition, which propels back to the conscious surface only the

* Professor Stefan Th. Gries, Ph.D., University of California, Santa Barbara, Department of Linguistics, USA - Santa Barbara, CA 93106-3100.

Journal of Institutional and Theoretical Economics 0, 1–22 – ISSN 0932-4569
DOI: 10.1628/jite-2026-0013 – © 2026 Mohr Siebeck

result of intuitive reasoning. Now, legal decision-making is rarely well described as being purely intuitive, so doctrine can come in and, arguably, guide the formation of the intuitive result.

One area where this can be exceedingly important is that of ordinary meaning in legal interpretation. The ordinary meaning doctrine is extremely important to many aspects of legal interpretation in the United States. It has been summarized, e.g., by

- Hobbs (2011, p. 328) as "words that are not defined by a statute be given their ordinary meaning by the interpreting court";
- (Eskridge, Slocum, and Gries, 2021, p. 1516f.) as "[focusing] on how an average reader – the typical member of the public – would understand the relevant language";
- Justice Holmes (1899, p. 417) famously opining that the interpreter's role is not to ask what the author meant to convey but instead to determine "what those words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used".

However, as Lee and Mouritsen (2018, p. 806) observe with regard to such ordinary meanings, "[t]ypically, this assessment is made at a gut level, on the basis of a judge's linguistic intuition", which can be highly problematic: Not only does the gut often not work so well, even more formally/scientifically, there is no uniformly accepted theory of ordinary meaning (see Lee and Mouritsen, 2018), and the combination of these two factors opens the door to both outright errors or, maybe worse, motivated reasoning (see Gries et al. (2022) for an example). As for the latter – a theory of ordinary meaning – it has frequently been argued that courts do not distinguish clearly between ordinary, common, plain, prototypical, natural, etc., meaning (see Aprill, 1998 or Solan, 2003 on the interchangeability of "plain meaning" and "ordinary meaning", see also Hobbs, 2011, p. 328f.), and some also (try to) keep those separate. Lee and Mouritsen (2018, p. 800) indicate that judges in fact often use *ordinary meaning* to refer to any point on this continuum:

possible → common → most frequent → exclusive

And, in fact, sometimes *prototypical meaning* is also added to the mix, as when Lee and Mouritsen (2018, p. 801) argue that prototypes map very well onto the notion of the "picture [of (a member of) a category] evoked in the common mind".

As for the former – gut/intuition – it is hard to imagine that, while intuition certainly plays some role, that (i) it would be desirable for intuition to play a central role given how it runs the risk of undermining the fair-notice function of the legal texts, and that (ii) intuition is not a reliable guide, especially because, while judges are experts in *legal* language ("judges are well aware of the ordinary meanings of contractual and statutory terms (Solan, 1993b; 1995)", see Hobbs, 2011, p. 329f), they are not experts in *ordinary* language. Such mismatches between expert and non-expert/laymen's judgments are generally well-known in linguistics, in particular when it comes to seemingly simple acceptability or grammaticality judgments (see Labov, 1975 or Schütze, 1993).

Some judges are of course aware that often more than their intuitions is needed for arguments regarding the ordinary meaning of some expression, so some turn to various "sources of evidence", first and foremost dictionaries. Slocum (2015, p. 21) summarizes previous research showing that, while The Supreme Court of the United States (SCOTUS)'s use of dictionaries was virtually non-existent before 1987, now up to 1/3 of statutory decisions cite dictionary definitions. Similarly, as per Mouritsen (2010), the frequency of *ambiguity* in SCOTUS decisions has risen from 12 per million words (pmw) (in 1969) to >100 pmw (1970–2005), and the frequency of *plain meaning* in SCOTUS decisions has risen from 4.2 pmw (in 1969) to 15.43 pmw (1970–2005, and 15.77 pmw for ordinary meaning).

While dictionary usage has become much more frequent over the last 50 years (see, e.g., Aprill, 1998; Hoffman, 2002; Solan, 1993a; 1993b; 1995; 2003; Weinstein, 2005), supplementing intuition with them comes with many problems, most of which reduce to the more general, in fact trivial, fact that describing ordinary meanings is not the purpose of dictionaries. Dictionaries mostly prescribe meaning/use and usually do not identify ordinary meaning – they often at most provide "a decent two-line approximation of the word's meaning" (Solan, 1993a, p. 53f) and are not intended to be used to determine the outer boundaries of a word's applicability. While this mismatch between the purpose and the application of dictionaries is already fatal, dictionary use in the legal domain comes with even more problems: First, dictionaries do not always use objective criteria for distinguishing senses; second, they are not error-free in the first place; third, judges have used them wrongly as when they rely on the order of sense although the orders of senses might be meaningless or be chosen to facilitate exposition or reflect diachrony; finally, dictionaries can falsely imbue an attempt at using them for ordinary meaning with an aura of "scientific precision", they can be selected inconsistently or, even worse, selected with an agenda ("haphazard[ly] and subjective[ly]") (Hobbs, 2011, p. 331).[1]

One famous example of dictionary abuse (by SCOTUS) is *Smith v. U.S.*. Smith was arrested after offering to trade a MAC-10 automatic weapon for cocaine and charged with drug-trafficking, but the indictment also alleged that he had knowingly used the MAC-10 during and in relation to a drug trafficking crime leading to a sentence enhancement (as per § 924(c)(1)). He was convicted of all charges, but appealed, arguing that the enhancement should only apply when the firearm is used *as a weapon*. SCOTUS rejected the final appeal in a remarkably contradictory way: SCOTUS first claimed to follow the ordinary meaning doctrine, but then continued to argue that the fact that some dictionaries provided very general definitions of *use*, according to which it is possible to use a gun for barter, supported the notion that the ordinary meaning of *using a firearm* was "to trade it for something else". However, as Justice Scalia already argued elsewhere (*Chisom v. Roemer* (1991, p. 410)), the court's job "is not to scavenge the world of English usage to discover whether there is any possible meaning" – "our job is to determine … the ordinary meaning." In other

[1] See also Aprill (1998), Brudney and Baum (2013), or Lee and Mouritsen (2018, p. 810f).

words, one can use "use a weapon" to mean "trade it in for …", but does anyone do that? Slocum and Gries (2017) find that, out of 159 cases of *use* + WEAPON NOUN, 0 mean "trade/barter", and out of an additional 159 cases of *use* + concrete DO, 0 mean "trade/barter".

An even more worrying example is *Muscarello v. U.S.*. Muscarello set out to sell marijuana with a handgun locked in the glove compartment of his truck (until the time of his arrest). SCOTUS argued that this situation was the ordinary meaning of *carry a gun* by

- using multiple dictionaries and forcing sense-ranking onto dictionary entries;
- relying on etymologies (implying that the ordinary meaning of *carry* is illuminated by arguing that *carry* goes back to Latin *carum* ("cart");
- consulting additional sources that imply that SCOTUS thinks ordinary meaning is approximated by the *King James Bible*, *Robinson Crusoe*, *Moby Dick*, *The Magnificent Seven*, *MASH*, and *Sesame Street* (!);
- pretending that judges are empirical scientists. Justice Breyer, for instance, searched computerized newspaper data bases (*NYT* database and a *U.S. News* database) "look[ing] for sentences in which the words *carry*, *vehicle*, and *weapon* […] all appear. We found thousands of such sentences, and random sampling suggests that many, perhaps more than 1/3, are sentences used to convey […] carrying of guns in a car." (Judge Posner used similarly bad Google searches in his opinion in *U.S. v. Costello* (2012).)

However, and especially with regard to the last bullet point, one cannot help but think (i) "what about the 2/3 that don't?" and (ii) "but this only checks for $p(carry + weapon + vehicle)$ when what is at issue is p("carry in vehicle" | *carry+weapon*)". And, indeed, proper empirical studies find that the meaning "carry$_{\text{on person}}$" massively outnumbers "carry$_{\text{as conveyance}}$" (see Mouritsen, 2010 and Goldfarb, 2017).

To improve on this, some "founded" the field of law and corpus linguistics – largely Solan, but recent developments have been particularly influenced by Lee and Mouritsen (2018). They followed the ideas by Breyer (and Posner), but

- use properly designed corpora to, often, represent ordinary meaning/usage;
- use corpus-linguistic analysis tools to obtain frequencies of (co-)occurrence (and later dispersion) as well as collocation (with a bit of association measures), and concordances;
- frame their research questions more usefully (as opposed to Breyer's).

Still, nearly all of this work involved an extensionalist/reference-based approach to meaning, in which concepts/categories are defined by "lists" of exemplars that instantiate a concept/category or are referred to. That means, in a sense, that

- *carry a gun* meaning "carry$_{\text{as conveyance}}$" is operationalized by the number of *vehicle* in collocations or concordance context;
- *interpreter* meaning "translation, but spoken" is based on how many collocates of *interpreter* involve speaking as opposed to writing;

– *vehicle* including airplanes is operationalized by the number of plane-related collocates used around *vehicle*.

1.2 Feature-Based Approaches and/or LLMs to the Rescue?

These initial proposals were promising but also come with several problems: For instance, what absences or presences of examples or collocates mean something relevant? The fact that *tire* is not a frequent collocate of *vehicle* does not mean that the prototypical vehicle – arguably a car – does not have tires. And, what if certain words are not attested in corpora because either (i) they are simply not talked about (e.g., for sampling reasons) or (ii) they refer to concepts not attested at corpus time or in the population a corpus is representative for? And, what if concepts and/or society change over time? It is in response to points of critique and questions like these that Gries, Slocum, and Tobia (2024) (as well as Egbert and Lee, 2024) propose an intensionalist approach to ordinary meaning. Such an approach is not based on examples but on "definitions" and/or categories/prototypes. How would one go about defining categories and especially prototypes?

According to the classical approach towards definitions and categorization, something is a member of a category if it exhibits all necessary (*only if* …) and then jointly sufficient conditions (*whenever* …) of a category. For instance, *only if* [someone is male]$_1$, [an adult]$_2$, [unmarried]$_3$, and [never married before]$_4$, and *whenever* 1 to 4 hold, that someone is a bachelor. According to the arguably cognitively more realistic prototype approach (see Rosch, 1978; Lakoff, 1987; Taylor 2004; 2011), categories are "held together" by a prototype and the prototype of a category C is an abstract entity that exhibits the features f_{1-n} that have the highest cue validity for category C. The cue validity of a feature f for a category C has been defined in two main ways:

– an early way (see Rosch, 1978), according to which cue validity is the conditional probability $p(C \mid f)$, i.e. the probability that something is a member of category C if it exhibits feature f;
– a later way (see Bates and MacWhinney, 1982; 1989; MacWhinney, 2005), according to which cue validity is the product of
  — cue reliability $p(C \mid f)$ (i.e. the above definition of cue validity) and
  — cue availability ($p(f)$).

That means a cue validity of a feature f for a category C is high if many/most/all instances/members of C have feature f and many/most/all non-instance/non-members of C do not have f. Consider the category C of "birds" and the feature f of "has beak": all birds have beaks and nearly all no-birds have no beaks (the few exceptions include platypodes, echidnea, …). If the distribution of these features would be the one represented as in Table 1, the early cue validity would be $100/102$ while the later one would be $100/102 \times 102/100$.

Another example famous in jurisprudence and legal scholarship is the one the present paper will look at empirically in more detail: the category of "vehicle", which features in applications like the following three. First, there is *McBoyle v. U.S.* (1931),

Table 1
Inputs to Two Cue Validities of *f* "Has Beak" for Category *C* "Bird"

	Category C "bird": yes	Category C "bird": no	Sum
Feature "has beak": yes	100	2	102
Feature "has beak": no	0	98	98
Sum	100	100	200

a case that involved the question of whether moving a stolen airplane across state lines violates the National Motor Vehicle Theft Act and, thus, what a vehicle is. (Recall from above that, in traditional extensionalist law and corpus linguistics approaches, an airplane is more likely to be considered a vehicle if *airplane* occurs around *vehicle* in corpus data (from the right register(s)/time(s)); see Lee and Mouritsen (2018)). Second and probably similarly famous is Hart's famous (1958) hypothetical: "No vehicles in the park!": What does *vehicles* refer to and how does general and specific situational context affect interpretation? Finally, there is *State Automobile Mutual Insurance Company v. Hoyle* (1992, before the North Carolina Court of Appeals), which centered on the question of whether a home insurance policy that mandates medical payments to third parties injured not with a motor vehicle apply when the injuries were caused by a motorized go-cart.

When it comes to defining categories or determining whether something is an instance of a category, large language models (LLMs) might prove to be an interesting alternative. Users can ask them for definitions of things including (i) legal definitions of things but also (ii) everyday – i.e. ordinary – definitions of things. More specifically even, users could ask them for the individually necessary and jointly sufficient conditions that feature in legal as well as everyday/ordinary definitions of things; users could even ask them for prototypes, which might be the best approximation of ordinary meaning, and some of these results seem amazingly useful for an intensionalist approach; in what follows, I provide the results of two relevant questions asked to two LLMs about the category of vehicle, Perplexity[2] and DeepSeek.[3]

1.2.1 Perplexity. In one application to the category of "vehicle", Perplexity was primed to provide "qualified subject matter expert" views, to "offer multiple viewpoints or solutions". When Perplexity was asked, "what is the definition of "vehicle" and what necessary and sufficient conditions might it involve?", it responded with definitions

– based on the California Vehicle Code Section 670;
– based on the Oxford English Dictionary;

[2] See <://www.perplexity.ai>.
[3] See <://chat.deepseek.com>.

– based on a general engineering perspective;

and the following necessary and sufficient conditions:

– device or constructed object (not a natural object);
– capability for movement (being moved or moving itself);
– intended purpose of transporting people, animals, or goods;
– not stationary by design (i.e., excluding buildings);

as well as with edge cases and counterexamples:

– bicycles: excluded by CA law if "moved exclusively by human power";
– trains: excluded by CA law if "used exclusively upon stationary rails or tracks";
– wheelbarrows: may or not be considered, very dependent on jurisdiction and context;
– drones: increasingly considered vehicles in aviation law.

Perplexity was then asked the following follow-up question: "I am defining the notion of *prototype of a category C* as 'an abstract entity – i.e. an entity that might not actually exist in the real world – that combines the features with the highest *cue validity* for the category *C*". The notion of *cue validity of a feature for a category C* in turn is defined as follows: "A feature F has a high cue validity for a category C if (i) most or all members of C have feature F and (ii) none or nearly no non-members of C do not have feature F'. Against this background, what are prototypical features of the category "vehicle" and what is the prototypical vehicle?".

Its response considered a car (sedan) to be the prototype and included the following features and their cue validities:

– transportation function: very high cue validity
  – prevalence in *C*: all vehicles have a transportation function;
  – prevalence outside of *C*: few non-vehicles have a transportation function (e.g., conveyor belts or elevators);
– mobility mechanism (wheels, tracks, propellers, and jet engines): high cue validity
  – prevalence in *C*: most vehicles have such a mechanism;
  – prevalence outside of in C: few non-vehicles have such a mechanism (e.g., wheeled office chairs, gliders);
– human operability: moderate cue validity
  – prevalence in *C*: most vehicles require a driver or pilot
  – prevalence outside of *C*: few non-vehicles require a driver/pilot (e.g., remote controlled toys, also consider drones and autonomous vehicles);
– enclosed passenger/cargo space: moderate cue validity
  – prevalence in *C*: common in cars, planes, ships;
  – prevalence outside of C: rare (e.g., shopping carts, motorcycles, bikes).

Regardless of whether one agrees with all the proposed features or whether everyone would consider the proposed features to indeed be the highest-cue validity features, something not likely even with a group of nine extremely similar individuals (SCOTUS Justices, who are similarly old, very similarly trained, have the same profession, …), it is clear that each feature has an at least fairly high cue validity and that a car is a very reasonable guess at the prototype of "vehicle".

1.2.2 DeepSeek. When DeepSeek (with DeepThink R1) was primed in the same way and asked the same question, it responded with a general definition that already hinted at a prototype: "a means of conveyance or transport, *typically* designed to carry people or goods from one place to another" (my emphasis), but then followed this up with three more focused definitions:

- based on the California Vehicle Code Section 670: a constructed device that is
  - capable of propulsion/movement
  - operable on a highway
  - transport function (people or property)
  - non-human powered
- based on the OED and a general engineering perspective based on the features
  - mobility (self-propelled or externally powered)
  - designed for transport (people, cargo, equipment)

It also provided edge cases and ambiguities that highlight well where ambiguity might arise:

- electric scooters, whose classification in California depends on their speed;
- autonomous robots, whose classification depends on whether they are designed for transport;
- wheelchairs, which may (!) only be vehicles if they are not human-powered.

Asked for prototypical features (based on cue validities) and a prototype, DeepSeek suggested cars, motorcycles, but also delivery drones as "approaching the prototype". As for prototypical features and their cue validities, DeepSeek offered the following:

- transportation function: high cue validity
  - all vehicles are designed to transport;
  - few non-vehicles share this primary purpose (e.g., furniture, appliances);
- mobility: high cue validity
  - vehicles are inherently mobile
  - non-vehicles (e.g., buildings, trees) lack this feature
- wheels or propulsion mechanism: cue validity: moderate (because there are vehicles like boats or hovercrafts)
  - most vehicles use wheels, tracks, rotors, or jets for movement;
  - non-vehicles that nonetheless have wheels (e.g., office chairs) lack transportation as their primary function;

- human/motor operation: moderate cue validity
  - most vehicles require a driver, pilot, or autonomous system;
  - non-vehicles may have motors but lack transport roles;
- structural frame/chassis: moderate cue validity
  - vehicles typically have a chassis to support loads and systems;
  - non-vehicles may have frames but no mobility (e.g. shelves).

Again it seems fair to say that many of these features and suggestions are useful and appear "thoughtful" and – maybe expectedly, certainly reassuringly – overlap with the results returned by Perplexity; the inclusion of delivery drones is somewhat unexpected, though.

1.2.3 Issues with applications of LLMs/AIs. While much of the LLM output is very useful – maybe especially the proposed features, as we will see below – there are also potential problems that make it unlikely that LLMs/AIs are the panacea one might hope for. One is that LLMs still often hallucinate, i.e. they can provide responses that are blatantly false, making up events or facts that are simply not true or never happened. For example, Magesh et al. (2025) discuss previous work that finds that "general-purpose LLMs hallucinate on legal queries on average between 58% and 82% of the time" (p. 3), and their own empirical evaluation of even LLMs augmented with legal information finds that

> Commercially-available RAG-based [*RAG* stands for *Retrieval-Augmented Generation*, STG] legal research tools still hallucinate. Over 1 in 6 of our queries caused Lexis+ AI and Ask Practical Law AI to respond with misleading or false information. Westlaw hallucinated substantially more—one-third of its responses contained a hallucination. (p. 9)

Discussing these results more qualitatively, they find that "[h]allucinations can be insidious" and that

> these systems continue to struggle with elementary legal comprehension: describing the holding of a case [...], distinguishing between legal actors (e.g., between the arguments of a litigant and the holding of the court), and respecting the hierarchy of legal authority. Identifying these misunderstandings often requires close analysis of cited sources. These vulnerabilities remain problematic for AI adoption in a profession that requires precision, clarity, and fidelity. (p. 10)

Maybe just as damagingly, the output of LLMs is not necessarily replicable, which surfaces in several different ways. First, even small variations in prompts can lead to different responses and, actually, even identical prompts can lead to different responses. On the one hand, this can be due to (default) settings of hyperparameters such as *temperature* or *top P* (see Peeperkorn et al., 2024). But even when those are set to 0 (for temperature) or identical values (for top P) across applications, results may still vary across applications, which is due to aspects of their technical

implementation: The calculations done when LLMs are computed involve some inherent randomness due to, for instance,

– how real numbers are used and rounded in numerical processing;[4]
– how computations are done in parallel and, thus, for instance in different orders, which also leads to floating point problems;[5]
– how computations algorithms use mixed precision training, which lowers replicable precision across mathematical operations.[6]

Given that legal scholars have argued against Google as a "legal corpus tool", it is not clear that such characteristics would make them consider LLMs acceptable options.

Another issue that can add to the above is concerned with the fact that, maybe especially in the U.S. American context, it is often necessary to adopt a historical perspective, where issues of dynamicity play a role (see Eskridge, Slocum, and Gries, 2021):

– societal dynamicity, e.g. when society or the world changes); for examples, objects referred to with a category may not exist at the time a law was passed (e.g., Segways did not exist in 1931) or were very different (e.g., when arms today are now very different from when the Second Amendment was passed)
– linguistic dynamicity: the meaning of a term may change considerably (e.g., what is now referred to as *gender* was, in the American English of the 1960s, usually referred to with the word *sex*);
– normative dynamicity: the legal or moral/societal context changed (e.g. when being gay is not viewed, in fact *judged*, as sexual perversion or deviance anymore, as it was in the 1960s.

I am not aware of studies of the question of how reliable LLMs would be when it comes to historical questions or applications. I would assume that the utility and reliability of LLM output would deteriorate exponentially with the distance of the relevant time from the time from which most of the LLM training material is from: With LLMs trained on contemporary data, results concerning the 1970s might still be quite good, results concerning the 1930s will probably be worse, and results concerning the founding era of the U.S. and its constitution would probably be quite bad.

One approach that might appear promising in how it addresses at least some of these points is the use of RAG, where an analyst provides an LLM with relevant material for analysis. For instance, in the case of a historical question, an analyst might provide textual material from the relevant time period to the LLM and then task it with relying on this material for the analysis. This is an attractive option but

[4] See floating point precision, <https://www.pamelatoman.net/blog/2023/08/nondeterminism-in-llms/>, Riach, 2019.
[5] See Shanmugavelu et al., 2024.
[6] <https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html>.

even with such an approach problems remain, and new ones are introduced: First, even with additional input of an RAG kind, an LLM's output will still be affected a lot but its original pre-RAG training input and parameters. Second, that of course also means that weaknesses such as biases in the training data that should not affect legal reasoning – as when LLMs think doctors are male and nurses are female, an probabilistically correct fact, but one that should likely not affect legal reasoning – can still affect LLMs' output. Relatedly, we have seen above that RAG does not eliminate the problem of hallucinations even in highly factual speech contexts such as legal applications.

Third, depending on one's application, RAG shifts the problem to one of sampling strategies. RAG could in theory be easy if the case is concerned with interpretation of some term in precedent cases, then the training material could just be all relevant previous cases (if this is possible sizewise; see immediately below). But what if the case is concerned with ordinary meaning? In that case, the additional training data required for RAG requires careful curation by someone with a background in, minimally, sampling design in the social sciences or, maybe ideally, a corpus linguist with extensive training and experience in corpus compilation. Relatedly, the size of training material that RAG might require could exceed what at least current architectures can handle. For instance, if the relevant issue is one of ordinary meaning at a certain point of time, then feeding all the corpus data one has for that time might not be possible – would be really be able to feed an LLM 100 million words of corpus data?!

Finally and more specifically for our context here, it is less than obvious how well and reliably we can assign weights to defining or prototype features (other than the simple ordinal labeling employed by the LLMs above), and it is not clear that LLMs are able to handle declarative negative knowledge – knowledge of what is *not* true – well enough for reliable application in legal contexts (recall the above results from Magesh et al., 2025). And, there are still some unresolved issues regarding LLM's potential copyright infringement issues that might get in the way of more widespread LLM adoption anyway.

All of the above makes one wonder whether there might be a kind of "compromise position", and the following section will propose one way to proceed.

2 *A Corpus-Linguistic Approach*

We have seen that

- extensionalist corpus-linguistic approaches to legal interpretation seemed promising initially, but ultimately run into problems that intensionalist approach seem to be able to avoid;
- LLMs can help with identifying features that are relevant to an intensionalist semantics kind of approach, but are plagued with replicability and likely lack a good way to deal with historical/diachronic data.

Maybe one can combine (i) the more pedestrian and lower-tech tools of corpus linguistics and (prototype) semantics, which come with replicability, flexibility regarding the weighting of features and input data with (ii) the higher-tech tools of generative AI tools, and the remainder of this section proposes one such approach. This approach is is based on large amounts data – like LLMs – and on the highly computational development of a statistical model/representation of such data – like LLMs but it is also based on lesser amounts of data and a less comprehensive model of those data. While that, all other things being equal, would make the alternative to be proposed seem worse than LLMs, it also comes with notable advantages, the most important of which are its replicability, its customizability, and its relative ease of being supervised by a human. In what follows, I briefly outline the approach by means of an example.

2.1 The Approach and a First (Synchronic) Example Application

The approach to be discussed here is a combination of the *method of distributional semantics* (potentially augmented by human ratings) and the theory of *intensional semantics* (from the perspective of prototype theory); we will look at the category of "vehicle" – specifically, we will look at how much different things "are" vehicles – and the approach here is a streamlining of, with suggestions for improvement of, the methods employed by Gries et al. (2024).

The first main ingredient for this kind of approach is a mass of textual data that can serve as a training corpus (much like LLMs are trained on extremely large amounts of internet data). This training corpus is then processed by a distributional-semantics kind of method. For quite some time, corpus data like that were processed on the basis of term-document matrices – i.e. matrices that state for each word form (in the rows) how often it occurred in each document in the corpus (in the columns) – the cells of which were then weighted by some statistics (such as association or dispersion info like *PMI* or *tf-idf*); see e.g. Deerwester et al. (1988). Such simple vector- and matrix decomposition approaches were then followed by more advanced word embeddings approaches such as shallow uni-directional neural network approaches such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington, Socher, and Manning, 2014; Bojanowski et al., 2016), only to be surpassed by deep contextual bidirectional-embeddings approaches like BERT (Vaswani et al., 2017; Devlin et al., 2018), which in turn led to the development of the kind of large language models like ChatGPT, Perplexity, Claude, or DeepSeek. The present example utilizes a by now actually computationally old-fashioned approach – GloVe– but this has the advantages of (i) being perfectly replicable (because GloVe models are trained before an application and, thus, static/replicable) and (ii) showing how even statistically much simpler but fast and easily available approaches are already quite powerful.

For *step 1* of the current proposal, I will use a pre-trained GloVe model, one of the largest available, which is the result of applying GloVe to 840 billion word tokens

from a crawled internet data;[7] this model is essentially a matrix of 2.2 million different word types each summarized in a 300-dimensional vector; a model like this can be imported into R (R Core Team, 2025) very simply. Note that, to not overcrowd this paper with code even more, the code presented here is abbreviated; see the appendix for information regarding the full code:

```
crawl_model <- read.vectors("glove.840B.300d.txt", vectors=300,
binary=FALSE)
```

Step 2 is to compute a submodel for the category of interest, i.e. "vehicle":

```
VEHICLE <- crawl_model[[c("vehicle", "vehicles", "Vehicle", "Vehicles"),
   average=TRUE]]
```

For *step 3*, we compute a sub-model for each feature we consider important for the definition of the category of interest; of course, these could be also be defined in a bottom-up way from `crawl_model`. For computational efficiency, we store them all in one data structure `features_as_models` and, for diagnostic reasons, add one feature ("courage") that should not be relevant to the category of interest:

```
features_as_models <- list(
            SPACE=crawl_model[[c("enclosed","passenger", …), average=T]],
      OPERABILITY=crawl_model[[c("operable","driver", …), average=T]],
            MOTOR=crawl_model[[c("motor","engine", …), average=T]],
   MOBILITYDEVICE=crawl_model[[c("device","mobility", …), average=T]],
        TRANSPORT=crawl_model[[c("person","passenger", …), average=T]],
          COURAGE=crawl_model[[c("courage","courageous", …), average=T]])
```

Step 4 consists of computing the cue validities of the features for the category; these will be operationalized on the basis of the cosine similarities of each feature model with the vehicle model:

```
cue_validities <- setNames(
   rep(NA, length(features_as_models)),
   names(features_as_models))
for (i in seq(features_as_models)) {
   cue_validities[i] <- cosineSimilarity(
      features_as_models[[i]], VEHICLE) }
cue_validities %>% round(4) %>% sort(decreasing=TRUE) # check the result
SPACE      TRANSPORT    MOTOR    OPERABILITY   MOBILITYDEVICE    COURAGE
0.5306     0.5005       0.4992   0.4024        0.3494            -0.0136
```

Cosine similarities can result in negative values, as is happening here, but since we will want to use `cue_validities` as weights to compute means, we make them all positive by adding to each the smallest negative cue validity plus a small positive constant:

```
if (any(cue_validities<0)) {
   cue_validities <- cue_validities + -min(cue_validities) + 0.01
}
cue_validities %>% round(4) %>% sort(decreasing=T) # check the result
SPACE      TRANSPORT    MOTOR    OPERABILITY   MOBILITYDEVICE    COURAGE
0.5542     0.5241       0.5227   0.4260        0.3729            0.0100
```

[7] Available at: <https://nlp.stanford.edu/data/glove.840B.300d.zip>.

Thus, `cue_validities` expresses how important each feature is to the category "vehicle"; reassuringly, we can already see that the value for "courage" is much much lower than that of all other, more realistic features.

In *step 5*, we compute a sub-model for each candidate term; for diagnostic reasons, we again add a few terms that this approach should reveal to be bad examples of vehicles:

```
candidates_as_models <- list(
          CAR=crawl_model[[c("car", "cars"), average=T]],
        TRUCK=crawl_model[[c("truck", "trucks"), average=T]],
     FORKLIFT=crawl_model[[c("forklift", "forklifts"), average=T]],
     …
        TABLE=crawl_model[[c("table", "tables"), average=T]],
        GLOVE=crawl_model[[c("glove", "gloves"), average=T]],
       ATTACK=crawl_model[[c("attack", "attacks"), average=T]])
```

Step 6 consists of computing how similar each candidate term (model) is to each feature (model):

```
candidates_2_features <- matrix(
   rep(NA, length(candidates_as_models)*length(features_as_models)),
   ncol=length(features_as_models), dimnames=...)))
for (i in seq(nrow(candidates_2_features))) {
   for (j in seq(ncol(candidates_2_features))) {
      candidates_2_features[i,j] <- cosineSimilarity(
        candidates_as_models[[i]], features_as_models[[j]]) }}
candidates_2_features %>% round(4) # check the result
```

```
            FEATURES
CANDIDATES    SPACE OPERABILITY  MOTOR MOBILITYDEVICE TRANSPORT COURAGE
  CAR        0.5407      0.4646 0.5965         0.3483    0.5316  0.0620
  TRUCK      0.6082      0.4581 0.5398         0.3076    0.5456  0.0477
  FORKLIFT   0.3340      0.2738 0.3418         0.2443    0.2982  0.0010
  …
  TABLE      0.3220      0.2779 0.1573         0.2585    0.2766 -0.0025
  GLOVE      0.2622      0.2063 0.1285         0.2717    0.1913  0.1328
  ATTACK     0.1797      0.1949 0.1635         0.2737    0.2790  0.3058
```

Thus, we now have a matrix `candidates_2_features` that states how similar each candidate term (in the rows) is to each feature (in the columns). Again, the interim results seem encouraging: The candidate *attack* is the only term to score highly on the decidedly non-vehicle feature "courage" while the candidate *car* scores very highly on the feature "motor".

Step 7 is the final one, in which we compute a vehicleness score for each candidate, which is the mean of each row of `candidates_2_features` weighted by `cue_validities`. That makes the vehicleness scores be based on how similar each candidate is to the features weighted by the importance of the features.

```
vehicleness_scores <- apply(candidates_2_features, MARGIN=1,
   FUN=weighted.mean, w=cue_validities)
sort(vehicleness_scores, decreasing=TRUE) # check the result
       CAR      TRUCK       BOAT      PLANE        BUS       BIKE
0.50561526 0.50438036 0.44434074 0.44115399 0.40615126 0.38358589
WHEELCHAIR    SCOOTER  MOTORBIKE      YACHT   FORKLIFT     ROCKET
0.35668861 0.35060663 0.31994364 0.30369373 0.30199557 0.28852022
     TABLE       CAGE     ATTACK      GLOVE      PANTS      ROACH
0.25744087 0.25489524 0.21554558 0.20882269 0.15473849 0.06472709
```

Even for such a simplistic case study, the results are encouraging because

- the top of the list are all and only all the things that a "reasonable average person" would probably view as at least reasonable candidates for vehicles, from *car* to *forklift*;
- the bottom of the list are all and only all the things that a "reasonable average person" would probably not consider vehicles, from *table* to *roach*;
- extremely uncontroversial, if not downright obviously prototypical, candidate terms like *car* and *truck* are indeed ranked highest.

2.2 The Approach and a Second (Diachronic) Example Application

The above was an application that presupposed that each candidate term or category – *car*, *truck*, etc. – is represented in the training corpus; this is because step 5 required computing a sub-model for each candidate term. But what if a candidate term is not represented in a training corpus? In such scenarios, which could easily arise in diachronic applications, one could proceed as outlined here on the basis of the question of how much a Segway is a vehicle according to American English in the 1950s, i.e. at a time when Hart's famous "No vehicles in the park!" sign might have been posted but Segways did not yet exist.

For this kind of application, *steps 1 to 4* are the same as before – the only difference is that the model is now not one that has been pre-generated and downloaded, now it is a model generated from the 1950s decade of the Corpus of Historical American English (I used 35 training passes, a context window of 4, a minimum word frequency of 3, and skipgrams). Thus, we again compute a model for the term of interest (*vehicle*), a list called `features_as_models` that contains a model for each feature we consider relevant to the definition of the category of "vehicle", and a vector `cue_validities` that expresses how important each feature is to that category.

Steps 5–6, i.e. the steps where we assess how similar each candidate term is to each feature, are where this approach is different from the previous one and there are different ways in which it can be implemented. A (much too) simple approach, but one that is didactically convenient in a proof-of-concept paper like the present one, is to enter the features manually into a matrix, which is shown here for a selection of presumably good and bad candidates for the category "vehicle". In this example, I am assuming, for instance, that

- a Segway has "somewhat of a designated driver/freight space" but not as much as a forklift, a Segway's "driver cabin" is more open than that of, say, a car; correspondingly, the dedicated and enclosed space of a car or a truck score highest;
- a table has a dedicated space for goods, but is not enclosed at all and tables score low on motor and mobility;
- the operability value of 0.1 for *attack* is supposed to reflect that an attack involves the semantic category of an agent, etc

```
           FEATURES
CANDIDATES SPACE OPERABILITY MOTOR MOBILITYDEVICE TRANSPORT COURAGE
  SEGWAY    0.67         1.0     1              1         1     0.0
  CAR       1.00         1.0     1              1         1     0.0
  TRUCK     1.00         1.0     1              1         1     0.0
  FORKLIFT  0.80         1.0     1              1         1     0.0
  PLANE     1.00         1.0     1              1         1     0.0
  TABLE     0.40         0.2     0              0         0     0.0
  GLOVE     0.00         0.0     0              0         0     0.0
  ATTACK    0.00         0.1     0              0         0     0.5
```

Thus, the matrix `candidates_2_features` again states how similar each candidate term is to each feature – just this time, we entered those manually. However, while I entered these manually here based on my own intuition – after all, this is a proof-of-concept – it is important to point out that these values could be expressed or derived in a variety of ways depending on need and the legal situation. First, the above scores are numeric values between 0 and 1, but it would of course be possible to just use 0 (for "no") and 1 (for "yes") as Boolean logical values.

Second, one can obtain such values in many other more objective ways. For instance, if a law (e.g., the National Motor Vehicle Theft Act of 1919) or some kind of agency (e.g. a state's Department of Motor Vehicles or a federal regulatory agency) had decided on a set of necessary conditions, the ab-/presence of all those features could be 0/1-coded for each candidate term as above. Similarly, if legal precedent determined that, in order to qualify as a vehicle, the "candidate device" must be traveling on land (as SCOTUS decided in McBoyle in 1931), a feature TRAVELONLAND could be added, for which *Segway*, *car*, *truck*, and *forklift* could be coded as 1 while the other terms could be coded as 0. More empirically, the ratings could be based on survey data (or even experimental data such as reaction times) obtained from the target audience for a specific statute or regulation. And if the target audience was ordinary speakers, the proverbial John Doe or Erika Mustermann / Otto Normalverbraucher, then one could try and develop these features vectors in a more bottom-up way, indeed maybe from some embeddings model.

The final step, *step 7*, would again replicate the previous approach:

```
vehicleness_scores <- apply(candidates_2_features, MARGIN=1,
   FUN=weighted.mean, w=cue_validities)
sort(vehicleness_scores, decreasing=TRUE) # check the result
     CAR    TRUCK    PLANE FORKLIFT   SEGWAY    TABLE   ATTACK    GLOVE
0.995389 0.995389 0.995389 0.951061 0.922248 0.121239 0.018596 0.000000
```

Again, these are encouraging results because

- the top *n* are all likely vehicular in nature, with the top two again arguably representing the prototype;
- the bottom *n* are all not vehicles at all;
- Segways are listed last of the vehicular items, but still with a value that makes them belong to the vehicular class rather than the others.

3 Evaluation and Conclusion

The present discussion could not discuss all possible ways in which the present proposals may be effective; for example, while this paper is a streamlined and revised version of Gries et al. (2024), I skipped a way to arrive at intensionalist definitions and decisions that is based on the manual annotation of concordance lines. Nevertheless, it is worth pointing out a variety of advantages of the proposed approach.

First and compared to the traditional approach to legal interpretation, while (corpus) linguists are not in the business of trying to deprive judges/justices of their constitutional tasks, even many legal scholars and practitioners agree that current practices can be shaky and suffer from motivated reasoning. Continuing to

- not even have a widely accepted definition of ordinary meaning or have one but not stick with it or deviate from it in counterintuitive ways (by claiming that the first meaning that comes to mind is not it);
- engage in likely ideologically-motivated dictionary shopping;
- pretend that dictionaries reveal ordinary meanings, that orders of senses are meaningful (even when dictionary makers state they don't), or that etymologies or literary pieces hundreds of years old inform current ordinary meaning;
- pretend that nine people who are very homogeneous in terms of training, profession, and everyday life but who are also extremely different from the ordinary people whose language comprehension they claim to be able to intuit;
- have judges untrained in social and empirical sciences butcher empirical studies;

does not instill trust into the fair-notice function that the ordinary meaning doctrine is supposed to protect in the legal system. Neither do programmatic claims that judges/justices will just "have to bone up on some basic linguistic methodology" (Devlin et al., 2018, p. 872) – as I have asked elsewhere: would Lee and Mouritsen (or legal scholars holding similar views) seek medical advice for their endocrinological disease from someone who just "boned up on" their endocrinology knowledge by doing a one-day workshop on corpus linguistics and the law? Somehow I don't think so, yet we are supposed to believe that a judge can do a quick corpus study or a quick "LLM analysis" in their chambers and do right by a defendant. And would the kind of analysis – in terms of size and expertise that is required – even be permitted, given the prohibition against extrajudicial investigation that prohibit judges from conducting new independent research?

Similarly untrustworthy and insufficient are flat assertions, e.g., by Scalia and Garner (2012) that their fair reading method solves all interpretive problems because it "requires aptitude in language, sound judgment, the suppression of personal preferences regarding the outcome, and, with older texts, historical linguistic research, plus it is also said to require an ability to comprehend the purpose of the text". The method may require that, yes, but (i) it is unclear how "aptitude in language" is operationalized and (ii) as the recent past has clearly shown, *requiring* the suppression of personal preferences is not the same as that actually *happening*.

Somewhat ironically, it is Scalia in particular who had sometimes had very sharp insights into ordinary meaning: In his dissent to *Smith v. U.S.* (1993), it was him who incisively criticizes the majority that did not "appear to grasp the distinction between how a word can be used and how it ordinarily is used" and that it fails to consider context in its interpretation: "[t]o use an instrumentality [e.g., "use a firearm"] ordinarily means to use it for its intended purpose" rather than as "an article of commerce". However, the very same Scalia (with Garner) defines "vehicle" as "sizable wheeled conveyance" without

– defining *sizable* and without explaining whether bicycles or tricycles count (freight trikes certainly are sizable wheeled conveyances and can be much more sizable than, say, Segways, which, according to Scalia and Garner, are excluded from parks);
– without explaining what this means for (i) vehicles using only continuous tracks as a means of propulsion (are tanks not vehicles?), (ii) vehicles using both wheels and continuous tracks (like certain kinds of snowmobiles), or (iii) things that have wheels but do not use them as their main means of propulsion (like airplanes) are vehicles or not.

Not to mention that language, society, and norms exhibit dynamicity, i.e. change continuously: firearms at the time of the Second Amendment differ from those we have today, the words *gender* and *sex* were used differently in the U.S. in the 1960s (when Title VII of the Civil Rights Act was passed) from 2019/2020 (when SCOTUS had to interpret it in the case *Bostock v. Clayton County*); see Eskridge, Slocum, and Gries (2021).

While technical, legal meaning is probably indeed safest when left in the role of expert judges, it does seem as if, with all due respect for judges' constitutional role, both the theoretical notion of ordinary meaning and its concrete operationalization are in desperate need of some improvement. The approach proposed here is certainly not ideal, but it has several advantages over previous methods or alternatives:

– given its use of language models trained on ordinary language or corpora consisting of ordinary language, its main empirical ingredient is incredibly much closer to ordinary meaning than that of a judge's linguistic mind that has been "tainted" by processing entirely unordinary language for the last 20–30 years;
– given the (simpler) math involved in embeddings, the present approach is perfectly replicable (different applications will yield different results) and much more versatile: assembling a training corpus and generating an embeddings kind of model (or even a bidirectional transformer like BERT) is orders of magnitude cheaper and faster that than building that same training corpus and then generating large language model;
– nearly all of the process of using embedding models like here can be made such that it avoids tinkering with it: everything that goes into the analysis (training data, features, weightings, but also survey ratings or experimental results of the kind discussed in Gries et al. (2024), …) can be obtained from sources

unrelated to the current case, which rules out the kind of, let's call it, "tinkering with" data/evidence such as just maintaining a definition that fits a judge's agenda, dictionary shopping, or the intentional distortion of empirical data (as when Judge Mizelle's court misrepresents corpus data in *Health Freedom Defense Fund, Inc. v. Biden* (2022), a decision later vacated in June 2023).

- finally, the approach proposed here avoids all the shortcomings of extensionalist approaches to legal interpretation: it does not rely on mere frequencies of occurrence or co-occurrence, which do not address the questions of (i) meaningful vs. meaningless absences and (ii) what to do when
  - words are not attested in a corpus because the concept does not exist at corpus time (see the case of *Segway* in the 1950s);
  - words are not attested in a corpus because the concept is not talked about (see Gales and Solan (2019)'s discussion of the very prototypical-looking bird species of blue pittas that is nevertheless never talked about in American English corpora);
  - the concept has massively changed since corpus time (see the difference of "arms" from 1791 to e.g., 2008, when *D.C. v. Heller* was decided in New York).

The most obvious next necessary steps would be to broaden and also validate the approach more in terms of the kinds of data used for it (e.g., concordance data, ratings, and better kinds of replicable models), the ways in which the approach is used (e.g. can we at least use LLMs for feature identification and their operationalization in embeddings/transformer models?), and in terms of the range of words to which it is applied. For example, the somewhat famous case of *Lozman v. City of Riviera Beach*, 568 U.S. 115 (2013) centered on the question of whether a floating home qualifies as a "vessel", a question that SCOTUS answered in the negative (because the floating home had no means of propulsion or steering); ultimately, one would of course also want to see whether abstract nouns can be dealt with with this approach, too. Hopefully, future work will be able to validate or improve our current, largely haphazard approach to ordinary meaning – it would be in the interest of anyone who feels that fair notice is an important judicial concept.

Appendix

An HTML report knitted from a Quarto document with all calculations and results is available at <https://www.stgries.info/research/2025_STG_EmbeddingsIntensionalSemantics_JITE.html>.

References

Aprill, Ellen P. (1998), "Dictionary Shopping in the Supreme Court," *Arizona State Law Journal*, 30, 275–336.

Bates, Elizabeth and Brian MacWhinney (1982), "Functionalist Approaches to Grammar," in: Eric Wanner and Lila R. Gleitman (eds.), *Language Acquisition: The State of the Art*, Cambridge University Press, New York, NY, pp. 173–218.

—, —, and Elizabeth Bates (1989), "Functionalism and the Competition Model," in: Brian MacWhinney and Elizabeth Bates (eds.), *The Crosslinguistic Study of Sentence Processing*, Cambridge University Press, New York, NY, pp. 3–73.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016), "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Brudney, James J. and Lawrence Baum (2013), "Oasis or Mirage: The Supreme Court's Thirst for Dictionaries in the Rehnquist and Roberts Eras," *William and Mary Law Review*, 55(2), 483–580.

Deerwester, Scott, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Laura Beck (1988), "Improving Information Retrieval with Latent Semantic Indexing," Proceedings of the 51st Annual Meeting of the American Society for Information Science, vol. 25, pp. 36–40.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805.

Egbert, Jesse and Thomas R. Lee (2024), "Prototype-by-Component Analysis: A Corpus-Based, Intensional Approach to Ordinary Meaning in Statutory Interpretation," *Applied Corpus Linguistics*, 4(1), 100078.

Eskridge Jr., William, Brian Slocum, and Stefan Gries (2021), "The Meaning of Sex: Dynamic Words, Novel Applications, and Original Public Meaning," *Michigan Law Review*, 119(7), 1503–1580.

Gales, Tammy and Lawrence M. Solan (2019), "Revisiting a Classic Problem in Statutory Interpretation: Is a Minister a Laborer," *Georgia State University Law Review*, 36(5), 491–533.

Goldfarb, Neil (2017), "A Lawyer's Introduction to Meaning in the Framework of Corpus Linguistics," *BYU Law Review*, 6, 1359–1417.

Gries, Stefan Th., Michael Kranzlein, Nathan Schneider, Brian G. Slocum, and Kevin Tobia (2022), "Unmasking Textualism: Linguistic Misunderstanding in the Transit Mask Order Case and Beyond," *Columbia Law Review*, 122(8), 192–213.

—, Brian G. Slocum, and Kevin Tobia (2024), "Corpus-Linguistic Approaches to Lexical Statutory Meaning: Extensionalist vs. Intensionalist Approaches," *Applied Corpus Linguistics*, 4(1), 100079.

Hobbs, Pamela. (2011), "Defining the Law: (Mis)Using the Dictionary to Decide Cases," *Discourse Studies*, 13(3), 327–347.

Hoffman, Craig. (2002), "Parse the Sentence First: Curbing the Urge to Resort to the Dictionary When Interpreting Legal Texts," *New York University Journal of Legislation and Public Policy*, 6, 401–438.

Labov, William. (1975), "Empirical Foundations of Linguistic Theory," in: Robert Austerlitz (ed.), *The Scope of American Linguistics*, The Peter de Ridder Press, Lisse, pp. 77–133.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, IL.

Lee, Thomas R. and Stephen C. Mouritsen (2018), "Judging Ordinary Meaning," *The Yale Law Journal*, 27, 788–879.

MacWhinney, Brian. (2005), "A Unified Model of Language Acquisition," in: Judith F. Kroll and Annette M.N. De Groot (eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*, Oxford University Press, Oxford, pp. 49–67.

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, et al. (2025), "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *Journal of Empirical Legal Studies*, 22(2), 1–27.

Mikolov, Tomas Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), "Distributed Representations of Words and Phrases and Their Compositionality," arXiv:1310.4546.

Mouritsen, Stephen C. (2010), "The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning," *BYU Law Review*, 2010(5), 1915–1978.

Peeperkorn, Max, Tom Kouwenhoven, Dan Brown, and Anna Jordanous (2024), "Is Temperature the Creativity Parameter of Large Language Models?" arXiv:2405.00492.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014), "GloVe: Global Vectors for Word Representation," Proceedings of the EMNLP 2014, ACL, Doha, Qatar, https://nlp.stanford.edu/projects/glove.

R Core Team (2025), "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Riach, Duncan (2019), "Determinism in Deep Learning," Developer.Nvidia.Com, https://developer.download.nvidia.com/video/gputechconf/gtc/2019/presentation/s9911-determinism-in-deep-learning.pdf, accessed January 27, 2025.

Rosch, Eleanor. (1978), "Principles of Categorization," in: Eleanor Rosch and Barbara Lloyd (eds.), *Cognition and Categorization*, Lawrence Erlbaum, Hillsdale, NJ, pp. 27–48.

Scalia, Antonin and Bryan A. Garner (2012), *Reading Law: The Interpretation of Legal Texts*, Thomson/West, St. Paul, MN.

Schütze, Carson T. (1993), *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*, The University of Chicago Press, Chicago, IL.

Shanmugavelu, Sanjif, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, et al. (2024), "Impacts of Floating-Point Non-Associativity on Reproducibility for HPC and Deep Learning Applications," arXiv:2408.05148.

Slocum, Brian G. (2015), *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation*, The University of Chicago Press, Chicago, IL.

Solan, Lawrence M (1993a), "When Judges Use the Dictionary," *American Speech*, 68(1), 50–57.

— (1993b), *The Language of Judges*, University of Chicago Press, Chicago, IL.

— (1995), "Judicial Decisions and Linguistic Analysis: Is There a Linguist in the Court," *Washington University Law Quarterly*, 73, 1069–1079.

— (2003), "Finding Ordinary Meaning in the Dictionary," Marlyn Robinson (ed.), Language and the Law: Proceedings of a Conference December 6-8, Tarleton Law Library, The University of Texas School of Law, Buffalo, NY, pp. 255–278.

Taylor, John R. (2004), *Linguistic Categorization: Prototypes in Linguistic Theory*, 3rd ed, Clarendon Press, Oxford.

— (2011), "Prototype Theory," in: Claudia Maienborn Klaus Von Heusinger and Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning*, Gruyter Mouton, Berlin and Boston, pp. 643–664.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. (2017), "Attention Is All You Need," arXiv:1706.03762.

Weinstein, Jason (2005), "Against Dictionaries: Using Analogical Reasoning to Achieve a More Restrained Textualism," *University of Michigan Journal of Law Reform*, 38(3), 649–681.

Stefan Th. Gries
Department of Linguistics
UC Santa Barbara & JLU Giessen
Santa Barbara, CA 93106-3100
United States
stgries@linguistics.ucsb.edu