

Data Driven Contract Interpretation: Discovering “Plain Meaning” Through Quantitative Methods

Author : Omri Ben-Shahar

Date : June 13, 2018

Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics* (Nov. 4, 2017), available at [SSRN](#).

Interpretation of contractual text may be the most important task courts perform in contract disputes. It is also the least predictable. Courts fall back on archaic canons of interpretation and employ poorly defined and spongy concepts for eliciting the meaning of words. They sometimes use textual approaches, and other times admit extrinsic evidence to understand the context. As a result, contract interpretation is erratic, and the resolution of contract disputes becomes complex and costly.

Despite murmurs of judicial skepticism and mountains of academic criticism, the most commonly used criterion in contract interpretation is the “Plain Meaning Rule”—the idea that if the language is clear and unambiguous courts should *not* consider any extrinsic evidence. But how to tell if a word is susceptible to a single plain meaning? Is it enough to look at dictionaries or to invoke judicial imagination to determine the unambiguous plain meaning?

In my own work and teachings, I have been advocating for a shift towards a data-driven search for plain meaning. My [recent article](#) with [Lior Strahilevitz](#) proposed one such empirical interpretive method: using large surveys. Now, a major new contribution to this timely enterprise of data-driven interpretation is being proposed by [Stephen Mouritsen](#). In an original and provocative article, Mouritsen introduces a method of interpretation based on empirical linguistics, and demonstrates—quite dramatically—the improvements it delivers relative to existing methods.

Consider the following example, taken from Mouritsen’s article. An insurance contract covers bodily injuries, but explicitly excludes injuries arising from participation in “any sports.” If the injury is a result of recreational *snorkeling*, does the exclusion apply? Is snorkeling a “sport”?

[A federal court said no](#). The judge looked at Webster’s Dictionary and found that the definition of “sport” is “rule-based athletic competition.” Since snorkeling is not governed by any traditional set of rules and it is not competitive, the judge concluded that it is not a “sport” and thus injuries occurring from it are not excluded from coverage under the contract.

That analysis, Mouritsen shows, is deeply flawed. The same dictionaries used by the court to define “sport” as a rule-based competition also provide a second meaning for “sport”: a “physical activity that gives enjoyment or recreation.” Snorkeling is surely a physical recreation! It turns out that the court’s own methodology to prove a single plain meaning—consulting leading dictionaries—supports each of the two opposing interpretations advocated by the parties. And when a text is susceptible to two plausible interpretations, the plain meaning rule could no longer be invoked—and should not have been relied on—to resolve the dispute.

But the exciting contribution of this article is not in showing that a word is susceptible to more than one meaning—this is old news. The breakthrough is in applying an empirical method known as [corpus linguistics](#) to choose the more appropriate meaning among the two competing dictionary definitions. The method Mouritsen applies does not require the usual messy investigation into the contract’s

surrounding context. It applies, instead, a quantitative and objective analysis to determine how the word is typically used in natural language.

The technical application of corpus linguistics to contract interpretation is quite simple, not much more exacting than learning how to use Westlaw. A digital search for the disputed word or phrase is made in a large *corpus*—a database of texts that represent the language used by the parties. An appropriate database for many contract disputes is the freely available [Corpus of Contemporary American English](#) (“COCA”). The search results can then be sorted according to the most common words that typically co-occur with the word in question. Each of these frequent accompanying words provide a qualitative sense of context. Since there are many such frequently co-occurring words, a simple quantitative test can then determine which meaning and context are more common.

What does the corpus linguistics method tell us about “sport”? Mouritsen searched COCA and found the common words that most co-occur with “sport”—words like *professional, teams, fans, pro, Olympic*. He looked at the 100 most common contexts in which the word “sport” co-occurs and found, strikingly, that in only one case its usage referred unambiguously to recreational activity (bungee jumping), whereas at least 50 contexts referred explicitly to “sport” as rule-based athletic competition, and many others strongly and unambiguously suggested the same. Apart from a small subset of contexts not related to either meaning, the result is an overwhelming quantitative prevalence of “sport” as rule-based competition, with only exceedingly infrequent use in the alternative meaning of physical recreation. “To the extent that our understanding of plain meaning has a frequency component,” Mouritsen sensibly points out, “we might conclude that the plain meaning of *sport* is *rule-based competition*.” A conclusion, he notes, that is not available through qualitative introspection.

Under the corpus linguistics method, snorkeling—which is not a rule-based competition—should not be interpreted as a “sport.” The court, in that case, got it right; but for the wrong reason. Had the court truly relied on dictionaries and then applied the “susceptible to more than one meaning” test for ambiguity, it would have concluded that “sport” is ambiguous and either demanded additional evidence or applied the contra-proferentem tie breaker. Instead, with corpus linguistics interpretation, the court would reach an unambiguous definitive interpretation.

Corpus linguistics, Mouritsen shows, can also help identify circumstances where a specialized or less common meaning is likely intended—again, without lengthy legal proceeding over the credibility of extrinsic evidence.

While corpus linguistics deploys rigorous tools and large databases to identify common meaning of language, it is of course not free from various subjective judgments. Which corpus to use? How to interpret the co-occurring words? But these determinations are done in an explicit and systematic manner, no longer fudged in the cognitive depths of judicial intuition. The time has come for courts to give data-driven interpretation methods their deserved attention.

Cite as: Omri Ben-Shahar, *Data Driven Contract Interpretation: Discovering “Plain Meaning” Through Quantitative Methods*, JOTWELL (June 13, 2018) (reviewing Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics* (Nov. 4, 2017), available at SSRN), <https://contracts.jotwell.com/data-driven-contract-interpretation-discovering-plain-meaning-through-quantitative-methods/>.