#### 2.8.1 Identifying textual predictors

A comprehensive literature review of behavioral economics research, particularly Cumulative Prospect Theory (CPT) [kahneman1979; tversky1992; tversky1974; tversky1981], revealed the four predictive dimensions of engaging words: representativeness, ease-of-use, affect, and distribution, referred henceforth as READ. These predictive dimensions, also called attributes, are derived from heuristics that have been empirically validated as simple means of predicting decision-making [czerlinski1999; gigerenzer1999]. These heuristics can be mapped onto information attributes that are situational, stable, and controllable based on the important dimensions of Attribution Theory, which explains how individuals collect data to arrive at causal judgments [kelley1967]. [@]

The following sections describe each predictor and its suggested operationalization.

##### Representativeness

When judging the representativeness of a new stimulus or event, individuals typically evaluate the degree of similarity between the stimulus or event and a certain standard or process [kahneman1972]. In accordance, the representativeness heuristic is the tendency to assess the similarity of objects and organize them based on their association, allowing for increased functionality, familiarity, findability, and fluency. Based on the work of Kahneman and Tversky, representativeness attributes or predictors can be described as the group of meanings and associations associated with a piece of information. A word is a basic unit for the expression of meaning, but the mapping between words and meanings is many-to-many; a word may have one meaning (synonymy) or many meanings (polysemy). Generally, individuals prefer one word when selecting from a set of synonyms and prefer one meaning of a polysemous word over others to convey a meaning. This means that although a word can have several interpretations, is likely to be the most representative.

The representativeness attribute can be operationalized using semantic relation analysis, which evaluates equivalency, hierarchy, and association, as is described in the Methods section.

##### Ease-of-use

Ease-of-use in terms of textual information asserts that users prefer texts they find highly readable based on the level of their processing (cognitive/perceptual) fluency. Both perceptual fluency and readability have been shown to be stimuli and determinants of decision-making, perception, and memory [alter2008; tversky1974]. Accessibility, the ease of processing information, has been specifically shown to influence engagement [dvir2018c]. The ease-of-processing heuristic is a fundamental heuristic in metacognition that guides, and biases judgments about memory. Various studies have shown that information that is easy to process is judged to have been learned well, providing evidence of the fundamental importance of the ease-of-processing heuristic [kornell2011].

Language fluency is one of a variety of terms used to characterize or measure language ability, often in conjunction with accuracy and complexity. Although there are no widely agreed-upon definitions or measures of language fluency, individuals are generally considered fluent if their use of language appears fluid, coherent, and paced appropriately. This relates to the availability heuristic, a mental shortcut that helps individuals understand the world by using information that they find easy to recall.

Calculating the complexity of a phrase in terms of word length (number of characters) and syllable number is one means of operationalizing and quantifying ease-of-processing heuristic. Another is its readability, of which several means of calculation exist, including the Flesch-Kincaid test.

##### Affect

In addition to its denotational or literal meaning, the emotional connotation or association that a word or phrase carries is critical in IE. An emotional connotation is frequently described as either positive or negative regarding its pleasing or displeasing emotional connection. Valence or hedonic tone is an affective quality referring to the intrinsic attractiveness or “goodness” (positive valence) or the averseness or “badness” (negative valence) of an individual, event, object, or situation [finucane2000]. There is now compelling evidence that every stimulus evokes an affective evaluation that may occur subconsciously [kahneman2002a]. The original affect heuristic explains why individuals often rely on their emotions rather than concrete information when making decisions. While using it allows them to reach a conclusion quickly and easily, doing so can also distort their thinking and lead them to make suboptimal choices.

Affective valence has been shown to influence decision-making and ultimately engagement [batra1986; commuri2008; finucane2000]. Based on the original affect heuristic, which has been widely accepted as a major general-purpose heuristic, the affect attribute proposes that the level of a word’s sentiment will have a significant impact on its level of engagement. Emotional valence can be operationalized by performing sentiment analysis, which aims at measuring, understanding, and quantifying the sentiment behind a text, e.g., whether it is positive, negative, or neutral, using NLP, computational linguistics, and text analysis [liu2012].

##### Distribution

The distribution attribute states that a term’s frequency or level of dispersion or distinctiveness influences its familiarity and fluency, which in turn impacts its representativeness. The frequency of a word is positively correlated with its familiarity, which is also an indicator of its cultural association and readability. This correlation is related to the recognition heuristic, which posits that if one of two objects is recognized and the other is not, the recognized object is likely to have greater value with respect to a criterion. Having been used as a model in the psychology of judgment and decision-making [tversky1992], the recognition heuristic relates to the availability heuristic, which assists individuals in understanding the world by using information easy to recall. For example, when comparing familiar and unfamiliar phrases, users have little recourse but to base their judgments on ease of retrieval or recognition [kahneman2002a]. In a series of experiments, participants were shown to rely on feelings of familiarity when comparing uncertain quantities, such as the relative size of two cities [gigerenzer1996]. This finding provides evidence that inferences are made based on functionality and familiarity, as users tend to prefer alternatives that are more recognizable and relevant.

Word frequency is an important factor in computational linguistics and NLP. Research has found that very frequent words are read and understood quickly and easily [brysbaert2011; savin1963], with highly frequent words perceived and interpreted correctly and more easily than infrequent words [savin1963]. In a study of the words used in the Internet Movie Database (IMDB), word frequency was shown to impact IE directly; specifically, high-frequency words were found to have a positive impact on IE [dvir2018f; dvir2019d]. The distribution attribute can be operationalized using TDM to measure situational significance (saliency or scarcity). This attribute can be defined therefore in terms of scarcity – the less scarce a word is, the higher its distribution.