Charles Marsh

Cigarette Helmets & Horse Wars

Towards a Better Understanding of Noun Compound Interpretability

Professor Christiane Fellbaum (Advisor)
Professor Srinivas Bangalore (Reader)

Department of Computer Science
Bachelor of Science in Engineering
April 2015

Contents

Abstract

The computational linguistics community has shown resurgent interest in the research of noun compounds, or sequences of nouns used to describe a single entity. Human judges frequently encounter noun compounds, be they familiar, like *coffee cup* and *park bench*, or unfamiliar, like *cigarette helmet* and *horse war*. Astoundingly, even these unfamiliar compounds are often interpretable with very little effort and in a manner that is widely agreeable to judges. This ease of interpretation is a testament to the productivity, generativity, and diversity of language in general and noun compounds in particular. However, it is clear that certain combinations of nouns would produce compounds that are not interpretable, or at least, incapable of being interpreted in a sensible manner. For example, devising a reasonable interpretation for the compound *pork plum* would be a daunting, if not impossible task. In this thesis, we test the limits of both human creativity and noun compound productivity, asking the question: "What makes a noun compound interpretable?" Though simple in formulation, this question has received little attention in prior research on compounds. Our analysis revolves around a series of experiments run on Amazon's Mechanical Turk platform in which human judges were asked to interpret and paraphrase binary and ternary noun compounds that had been generated 'at random' using an algorithmic process. Throughout this thesis, we analyze the results of these experiments to construct a more complete theory of noun compound interpretability, demonstrating the usefulness of semantic and lexical similarity-based comparisons to familiar compounds in determining the degree to which a new, unfamiliar compound is itself interpretable, as well as the deep and even intrinsic link between the acts of paraphrasing and interpretation.

1 Introduction

Over the past few years, the Natural Language Processing (NLP) community has shown resurgent interest in the analysis of *noun compounds*, or "long sequences of nouns acting as a single noun", such as *coffee cup*, *steel knife*, and *shoe sale* [27]. Much of this interest has stemmed from the incredible productivity, generativity, and diversity of these structures, qualities which makes the noun compound a fascinating element of human language. For example, even with the simple wildcard pattern ($*$ *sale*), we can generate such noun compounds as *shoe sale*, *baby sale*, *fire sale*, and so forth, each of which involves a different semantic relationship between the two nouns: a *shoe sale* would typically be paraphrased as "a sale of shoes"; a *baby sale*, as "a sale of clothes for babies"; and a *fire sale*, as "a rapid sale of goods", with this latter example representing an idiomatic expression.

Along with this incredible productivity, noun compounds can also be arbitrarily long–for example, *lung cancer treatment* is a length-3 compound–and can be either compositional (i.e., with a meaning derived through some combination of the meanings of its components) or non-compositional, as evidenced by the *fire sale* example, in which the meaning of the word "fire" plays no role. But what is perhaps most fascinating about noun compounds is the relative ease with which they are interpretable and understandable by humans: even with unfamiliar compounds, humans are able to come up with reasonable interpretations with which other

judges often agree. Returning to the ($*$ *sale*) example: whether or not one has encountered the compound *apple sale* in the past, they will likely settle upon an interpretation along the lines of "a sale of apples" (and do so with little difficulty or hesitation), which many judges would find reasonable.

Given these remarkable properties, it should come as no surprise that the analysis of noun compounds has extensive applications in language processing. For example, statistical machine translation can be greatly aided by generating accurate paraphrases for these short noun compounds, as seen in the work of Nakov and Hearst [27], where the authors expand compounds like *apple juice* to phrases like "juice that is made from apples", leading to more accurate translation. Similarly, question answering systems must disambiguate noun compounds in order to provide correct answers with any degree of certainty [1]. Returning to the example from earlier: a question answering system that treated a *baby sale* as "a sale of babies" would be useless to consumers.

1.1 Motivation

While noun compounds are an incredibly diverse and productive linguistic structure, much of the existing research on compounds has focused on two primary tasks:

1. Developing taxonomies through which to classify noun compounds according to the semantic relationships between their constituent components.

2. Developing techniques through which to paraphrase noun compounds.

However, academics have paid little attention to the somewhat deeper question of whether a given noun compound has *any* meaning at all. In other words: given a noun compound, can a human judge come up with a valid interpretation? Furthermore, is there one interpretation on which human judges would agree? Or multiple?

Put differently: is there a limit to the productivity of noun compounds?

In a sense, these are questions that test the limits of human creativity: as discussed in Section 1, noun compounds are incredibly productive and diverse–humans are constantly encountering and creating new compounds, often agreeing on their semantic meaning with impressive consistency–yet surely there are limits to what humans would consider interpretable.

For example, the compound *summer dispute* could be reasonably interpreted as "a dispute that takes place over the summer" or, just as reasonably, "a dispute about events in the summer". On the other hand, it is difficult to provide any such valid interpretation for the compound *pork wall*–any suggested interpretation would verge on or venture deeply into the realm of nonsense.

In this thesis, we explore the questions presented above in an unprecedented attempt to develop a more complete understanding of noun compound interpretability. Our efforts focus on pushing the limits of interpretability; in doing so, we employ a variety of semantic and lexical similarity metrics, including those derived from WordNet [10], to demonstrate

their applicability to noun compound understanding, and explore such tasks as training a machine learning classifier to judge compound interpretability and clustering compounds based on the grammatical structure of user-submitted paraphrases.

1.2 Outline

This thesis is structured as follows: Section 2 is used to lay out the necessary background information, including prior research and crucial terminology; in Section 3, we present our initial hypotheses which will be explored and expanded upon throughout the remainder of the thesis; in Section 4, we describe the design of our experiments, which were conducted using the Amazon Mechanical Turk platform; we then present the initial results from said experiments in Section 5; this is followed by thorough analysis of said results in Section 6; next, in Section 7, we compare the interpretability of semantically similar *peer* compounds; we then examine the interpretability of *ternary* compounds, or those composed of three words, in Section 8. We conclude with a discussion of the results in Section 9 and possible extensions in Section 10.

2 Background

We begin by exploring the current understanding of noun compounds, including relevant prior research, before diving into paper-specific terminology and other details that will be crucial in parsing the remainder of this thesis.

2.1 A Primer on Noun Compounds

If a compound consists of just two words, like *olive oil*, it is referred to as a *binary compound*. Similarly, a compound consisting of three words, like *olive oil bottle*, is referred to as a *ternary compound*.

As a simplifying assumption, noun compound research tends to focus on the analysis of binary compounds. This tendency is exemplified by the datasets of Hermann et al. [13], Kim and Baldwin [17], Nakov and Hearst [27], Ó Séaghdha and Copestake [31], Peñas and Ovchinnikova [34], and Tratz and Hovy [41], all of which contain exclusively binary compounds. While the majority of this paper focuses on binary compounds, we extend our analysis to ternary compounds in Section 8.

The Head-Modifier Principle

In English, Binary compounds typically involve the first element (an attributive noun) modifying the second element. As such, for a binary compound, the first word is referred to as the *modifier*, while the second word is the *head* [27].[1]

[1] This is true of some, but not all languages. For example, in French, the head is typically on the left [30].

In the case of *olive oil*, then, *olive* is the modifier and *oil* is the head, as the word *olive* is describing the type of *oil*. Similarly, for *coffee cup*, *coffee* is the modifier and *cup* the head, as the word *coffee* is describing the type or purpose of the *cup*.

This principle, which we refer to as the Head-Modifier Principle, is assumed to hold true throughout this thesis.[2] That is, when we refer to the modifier or head of a compound, we are implicitly referring to its left and right components, respectively.

Note that the Head-Modifier Principle

Branching

For compounds composed of more than two components, such as ternary compounds, the Head-Modifier Principle can be applied recursively. Typically, such compounds are described in terms of *branching*, a phrase which refers to the way in which the words are grouped together during parsing [20].

For example, we could define the ternary compound *olive oil bottle* to be "a bottle of olive oil". The definition would mark *olive oil bottle* as a *left-branching* compound. We could then bracket this interpretation like so: *[[olive oil] bottle]*. This bracketing syntax preserves the mental grouping of the words in the compound, as we have defined *olive oil bottle* such that the words *olive oil* are modifying the word *bottle*. In this case, *olive oil*, a sub-compound, is the modifier, and *bottle* is the head. Going one level deeper, we could then say that *olive* is the modifier and *oil* the head of the sub-compound *olive oil*.

As an alternative example, *chocolate birthday cake* is a *right-branching* compound when defined as "a birthday cake made of chocolate". Thus, it would be represented as *[chocolate [birthday cake]]*, with *chocolate* the modifier and *birthday cake*, a sub-compound, the head.

This representation can be applied recursively to compounds of arbitrary length. For example, *[[chocolate [[birthday party] cake]] obsession]* is a length-5 compound with a variety of left- and right-branching sub-compounds.

These designations (left- and right-branching) will be important when analyzing the interpretability of ternary compounds in Section 8.

The Principle of Compositionality

There are a number of theories as to how the human brain processes new or unfamiliar noun compounds, the most popular of which is known as the Principle of Compositionality.

Definition (Principle of Compositionality)**.** *The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined [33].*

In the context of noun compounds, this principle states that the act of parsing a new compound involves parsing its head and modifier separately, and then finding some means by which to combine them. For example, to interpret the compound *brick wall*, a human

[2]The English language does contain some compounds that do not follow the Head-Modifier Principle, like the 'exocentric' compound *pickpocket*. However, these compounds are relatively rare and typically non-productive [3]. As such, they are ignored throughout this thesis.

judge might first parse *brick* and *wall* independently, and then reason that the *wall* could be made of *brick*, thus developing an interpretation through the combination of two separate definitions. Note that this process implicitly involves the act of disambiguation, as the interpreter must decide on the optimal sense in which each word should be used (e.g., in the compound *apple juice*, the judge must decide that *apple* is best interpreted as a fruit, rather than, say, a computer brand).

The Principle of Compositionality is in opposition to, say, a theory of understanding that requires the brain to store every compound it has seen and subsequently search for a similar compound with which it is familiar when parsing new or unfamiliar compounds.

Given the straightforward nature of this Principle, we find it particularly appealing, especially as it relates to noun compounds; in this thesis, we generally assume it to hold true.

2.2 Prior Research

As mentioned in Section 1.1, there is little to no existing research investigating the question of noun compound interpretability. Instead, academics have typically focused on constructing classification taxonomies for the various semantic relationships between words in a compound, and techniques for automated paraphrase generation.

The Syntax and Semantics of Complex Nominals

As mentioned previously, many academics have focused on constructing classification taxonomies for the semantic relationships present in noun compounds. A discussion of the history of these taxonomies will be useful in tracing the history of noun compounds, as well as highlighting their distinctive qualities.

The first such taxonomy was introduced in 1978 with the publishing of Levi's seminal book, *The Syntax and Semantics of Complex Nominals*. Levi proposed a flat taxonomy of nine different semantic relationships: `cause`, `have`, `make`, `use`, `be`, `in`, `for`, `from`, `about`. According to Levi, these nine classifications cover a broad swath of noun compounds. For example, the compound *honey bee* would be classified as `make`, as in, "the bee `makes` honey." Similarly, the compound *tear gas* would be classified as `cause`, as in, "gas that `causes` tears."

But as Newmeyer noted shortly after publication, these classifications fail to accurately capture the 'meaning' of a compound. Looking again at the *tear gas* example: tear gas is not just a gas that causes tears. As Newmeyer remarks: "if it has a paraphrasable meaning at all, it is 'gas so-called because one of its properties is to cause tears'" [29].

Levi's response is to claim that these classifications are not meant to represent the 'meaning' of a compound, but rather, the realm of interpretations that one might use when encountering a new, unseen compound. The distinction is subtle but significant. For example, if one had never encountered the compound *tear gas*, they would likely interpret it as a 'gas that `causes` tears'. So although `cause` may not accurately capture the meaning of *tear gas*, it could capture a reasonable interpretation.

This makes Levi's schema more of a theoretical construct than a practical tool. Indeed, there are other difficulties when looking at the proposed schema as a 'tool'. For one, when we confine ourselves to a set of just nine semantic relationships, a great deal of ambiguity emerges, as the relationships included in the set become overly broad–and necessarily so, given the incredible diversity of compounds that they must cover. Using the *tear gas* example again, one could make a reasonable claim for assigning it to the `cause`, `make`, or `for` categories.

Advanced Taxonomies

In response to these difficulties, academics have continued to iterate on the idea of constructing fixed taxonomies of semantic relations for noun compounds. Over time, these taxonomies have grown in size, from the 13 semantic relations of Vanderwende [43], to the 20 relations of Barker and Szpakowicz [2], to the 30 relations of Nastase and Hearst [28], and so forth.

The most ambitious and comprehensive effort to enumerate an explicit list of relation classifiers can be found in Tratz and Hovy [41], which provides 43 compound relations in an attempt to "start fresh and build a new taxonomy" given the heterogeneity of previous attempts.

However, Tratz and Hovy readily admit to the existence of "an unbounded number of relations". Thus, their taxonomy merely aims to cover the "vast majority" of noun compounds, as the authors admit to the potential impossibility of comprehensive coverage.

With many of these taxonomies, the difficulty is in striking a balance: too few categories, and the classifications are overly broad and ambiguous, leading to schema that are not useful in practice; too many categories, and human judges fail to agree on the correct classifications. In the case of Tratz and Hovy [41], for example, their results, while respectable, did not represent a substantial statistical improvement over existing taxonomies in terms of co-agreement, the level of consensus among judges as to which designation best captures the relationship exhibited by the compound.

An Unbounded Set of Relations

Recently, noun compound research has taken an interesting turn, in part due to the work of Nakov and Hearst [27]. Instead of focusing on the construction of fixed taxonomies, academics have instead developed techniques for classifying semantic relationships in noun compounds through the use of *verbs*. Specifically, rather than describing a head-modifier pair as `from` or `be`, the goal is to describe in terms of a paraphrasing verb that accurately captures the relationship.

As an example: while Levi's schema might label *cancer doctor* as `for` (i.e., "a doctor `for` cancer"), using verbs gives us the flexibility to label it as `specialize` or `treat`. This approach leads to more accurate classifications and, as an added bonus, more faithful paraphrases, e.g., "a cancer doctor is a doctor that `specializes in` cancer". In effect, the use of paraphrasing verbs has two advantages: firstly, it allows for the use of an unbounded set

of relations (assuming that there are an infinite number of verbs); and secondly, it allows for the creation of paraphrases that are immediately useful.

In Nakov and Hearst [27], the authors develop a search-engine based technique for producing verbs that describe the semantic relation underpinning a noun compound. Their algorithms beat the baseline on a number of NLP tasks, suggesting that verbs are "the single most important feature for predicting semantic relations", among those considered in the article.

This development is further evidence of the supreme productivity and diversity of compounds, qualities that evidently cannot be captured by a fixed taxonomy. As far as this development is relevant to the present study: as in Nakov and Hearst [27], we use verbs and prepositions, through use of the relative clause, to paraphrase noun compounds, which stems from a belief that these expressive techniques are necessary for capturing the realm of possible interpretations.

Lexical & Semantic Similarity

Finally, we briefly review the use of lexical and semantic similarity techniques in classifying and paraphrasing noun compounds.

Along with creating classification taxonomies, researchers have also focused on techniques for automated classification of compounds based on different sets of semantic relations. In Kim and Baldwin [17], for example, the authors develop techniques for automatic noun compound classification based on WordNet similarity metrics. In particular, Kim and Baldwin focus on classifying compounds based on the semantic relations between their constituent components. Their results suggest that these metrics, which are based on semantic similarity, are helpful in categorizing noun compounds. While the questions they address are very different from those in this thesis, their techniques are similar.

Ó Séaghdha and Copestake [32] make use of lexical and relational similarity features, focusing on co-occurrence and other corpus-based techniques. As above, their results demonstrate that these feature choices are useful in capturing the semantic meaning of noun compounds. Thus, we make use of lexical features in our analysis as well.

The usefulness of these features in prior work suggests that speakers likely produce and interpret new compounds based on a process of analogy and comparison, given the intimate links between these processes and the concept of semantic similarity.

2.3 WordNet

As much of the analysis and terminology that follows relies on a basic understanding of WordNet, we include a brief description of its fundamental properties and principles for completeness, based on Fellbaum [10].

WordNet is a lexical database of the English language that covers nouns, verbs, adjectives, and adverbs.[3] WordNet's core atomic unit is the *synset*, which can be thought of as an

[3]For the purposes of this paper, the noun graph is the only graph of importance.

unordered set of synonyms representing a single, distinct concept, and usually includes a brief *gloss*, or definition, along with some examples.

In WordNet, English-language words can map to multiple synsets if they're used in multiple different senses. For example, the synset *paper.n.1* represents paper the substance, while the synset *paper.n.5* represents the idea of an academic paper.

Synsets are connected in a graph based on the concept of hyperonymy. Specifically, in WordNet, a connection from one synset to another going down in the graph represents a more general synset becoming increasingly specific. In particular, the *hypernym* of a given synset is the parent synset, which, in theory, represents a more general concept of which the initial synset is an instance. Similarly, the *hyponyms* of a given synset are its children, which should be even more specific. In this way, going deeper in the graph leads to more specific synsets, while going higher (i.e., towards the root) leads to those that are more abstract.

For example, the synset *apple.n.1* has the gloss "fruit with red or yellow or green skin and sweet to tart crisp whitish flesh". Its hypernym, *edible_fruit.n.1*, has the gloss "edible reproductive body of a seed plant especially one having sweet flesh", representing the more abstract idea of fruit that can be eaten, of which an apple is an instance. Meanwhile, *apple.n.1*'s hyponyms include *cooking_apple.n.1*, which has the gloss "an apple used primarily in cooking for pies and applesauce etc.", representing a specific type or usage of an apple.

2.4 Terminology

In this section, we begin to introduce some of the methodology underlying this thesis, along with the terminology that will be necessary to parse the remainder of the report.

Compound Types

For the remainder of the paper, we rely on the following terminology: *attested compounds* are noun compounds that are found in either the Kim and Baldwin [17], Nakov and Hearst [27], Ó Séaghdha and Copestake [31], or Tratz and Hovy [41]. As each of these datasets sourced its compounds from substantial corpora of English text, attested compounds are assumed to be easily interpretable (and often familiar) to human judges. To provide a sense of scale, these datasets combined to produce 20,710 distinct binary compounds.

At the opposite end of the spectrum, *generated compounds* are those which were constructed algorithmically, as described in Section 4.1. Generated compounds are unattested in that they are purposefully not present in the aforementioned datasets; this leaves open the possibility of a generated compound being difficult or impossible to interpret. In brief, generated compounds are created by combining two unrelated words, regardless of whether or not they pair well together. As such, generated compounds are likely to be new to human judges given the sheer immensity of the domain and the random nature of their construction; they are intentionally high entropy. The relationship between attested and generated compounds is depicted in Figure 1; this process is described in more detail in Section 4.1 below.

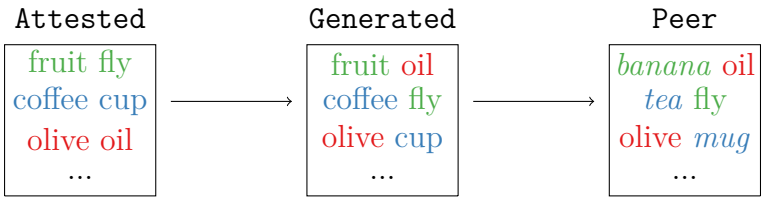


Figure 1: Attested compounds, on the left, are familiar and easy to interpret. Generated compounds, in the middle, are often new to human judges. Peer compounds, on the right, are created by mutating a constituent component of a generated compound using WordNet synsets.

An instance of a generated compound is said to be *annotated* if multiple human judges have indicated whether or not they were able to interpret the compound and, if possible, provided paraphrases based on their interpretations. Later, in Section 6.1, we describe a process by which these paraphrases were used to determine the WordNet synsets that best represent the sense in which a noun compound's head and modifier were employed. The determination of these synsets is another part of this annotation process.

Finally, *peer compounds* are those created by taking an annotated, generated compound and modifying either its modifier or head via a systematic walk through WordNet. Recall that annotated compounds, by definition, include the WordNet synsets that best capture their interpretation. To construct a peer, then, we use these synsets to replace a compound's head or modifier with a new word extracted from a nearby WordNet synset. Peers thus exist relative to the generated compound from which they were constructed. As a concrete example, if the word *party* in the annotated compound *party cake* were found to be best represented by the modifier synset *party.n.2*, then by taking the hyponym of this synset (*wedding.n.3*), we could produce the peer *wedding cake*. In this case, *wedding cake* would be the peer compound, while *party cake* would be its *root*.

In this thesis, we consider four such peer relationships, as defined in Table 1 and illustrated in Figure 2:

Name	Relationship to `Root`
`Child`	Hyponym
`Sibling`	Co-hypernym
`Uncle`	Hypernym of co-hypernym
`Nephew`	Hyponym of co-hypernym

Table 1: Precise WordNet relations and their aliases.

Interpretability Classifications

We define three categories of interpretability for a given noun compound:

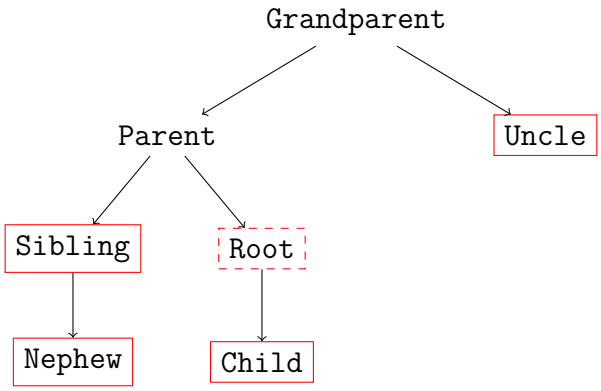


Figure 2: WordNet relationships from a given root synset `Root`.

- A compound is said to be interpretable with *No difficulty* if a human judge can easily decide on a meaning for the compound. Compounds like *olive oil* and *coffee cup* are interpretable with *No difficulty*.
- A compound is said to be interpretable with *Minor difficulty* if a human judge can come up with a reasonable but awkward meaning. Compounds like *deficiency committee* and *horse structure* are interpretable with *Minor difficulty*, according to human judges.
- A compound is said to be *Meaningless* if a human judge cannot come up with a reasonable meaning for the compound or if it makes no sense to said judge. Compounds like *rice eye* and *daisy baby* are *Meaningless*, according to human judges. Note that it is occasionally *possible* to come up with a paraphrase for a *Meaningless* compound, but that paraphrase should be considered nonsensical by the judge.

These labels are utilized in the experiments described in Section 4 as well as the analysis that follows.

3 Hypotheses

With the key terms established, we now define the hypotheses underlying our investigation of interpretability. These are provided as a series of claims, each of which will be examined in detail throughout this thesis:

H1. Given the creativity and productivity of noun compounds as a linguistic structure, only a small minority of compounds should be uninterpretable. In particular, the majority of compounds should be classified as interpretable with *No difficulty* or *Minor difficulty*.

H2. Compounds that are easier to interpret (i.e., interpretable with *No difficulty* rather than *Minor difficulty*) should tend towards fewer interpretations (as in, less variety). In particular, compounds that are easy to interpret should be less ambiguous (than

those interpretable with *Minor difficulty*) and human judges should therefore exhibit less variety and higher inter-rater agreement in their interpretations.

H3. Given the usefulness of WordNet-based (and other) semantic similarity metrics, as seen in the work of Kim and Baldwin [17], WordNet-based analysis should be effective for understanding and even classifying compounds by ease of interpretability. In particular, WordNet-based comparisons between attested and generated compounds should provide insights into compound interpretability. This assumes that human judges create and understand new compounds through a process of analogy based on knowledge of existing, frequently occurring compounds. Validation of this hypothesis would further evidence the assumption.

H4. The contribution of the head and modifier in comparing compounds to their attested variants should be roughly equal, in agreement with Kim and Baldwin [17]. In particular, distinct machine learning systems designed around usage of the head and modifier for comparisons between generated and attested synsets should exhibit similar accuracy.

H5. Human-provided paraphrases for compounds interpretable with *Minor difficulty* should require more and a greater diversity of tokens (i.e., words) due to the (evidently) unusual nature of the compound, which, in-turn, requires that a greater volume of information be expressed through the paraphrase. In particular, paraphrases for *Minor difficulty* compounds should use fewer prepositions and more verbs, as verbs are considered more expressive [27], and contain more and a greater diversity of tokens.

H6. Peer compounds should exhibit patterns in interpretability vis-à-vis their roots depending on the relationship between them. In particular, peers created by the `Sibling` relation should be more likely to have the same interpretability label as its root and, more generally, the more similar a peer to its root, the closer it should be in interpretability.

Many of these hypotheses will be expanded upon throughout the paper, and each will be assessed for validity. By presenting them at this early juncture, we hope to make our intentions and beliefs clear before presenting or analyzing the relevant data.

4 Experimental Design

At the core of this thesis is a series of experiments run on the Amazon Mechanical Turk (AMT) platform, a website which allows experimenters (also known as *requesters*) to design and create small experiments, known as Human Intelligence Tasks (HITs), to be completed by human workers (also known as *Turkers*) [6].

In total, we ran three rounds of experiments. In this section, we focus on the first round, the goal of which was to provide sufficient data so as to make observations validating or invalidating the hypotheses outlined in Section 3. The second round of experiments, which

focused on peer compounds, is described in Section 7, with the third round, which focused on ternary compounds, described in Section 8.

At a high level, this first round of experiments asked human judges to interpret generated compounds using the template described in Section 4.2. These judgments were then used to address the hypotheses from Section 3.

4.1 Data Generation

In this first round of experiments, 250 noun compounds were generated through the following process:

1. A set of base compounds was computed by taking the union of the Kim and Baldwin [17], Nakov and Hearst [27], Ó Séaghdha and Copestake [31], and Tratz and Hovy [41] datasets. Note that this set of compounds is equivalent to that of the attested compounds defined in Section 2.4.

2. Each compound was divided into a modifier and a head, which led to two sets, the first consisting of all possible modifiers and the second, all possible heads.

3. A randomly chosen modifier and a randomly chosen head were then concatenated. The resulting compound was discarded if: (1) it was attested (i.e., it was present in any of the existing datasets), or (2) it had already been generated. This step was repeated until 250 distinct compounds had been produced.

The resulting compounds, which consisted of a randomly paired head and modifier, represent the *generated compounds* defined in Section 2.4; this process is depicted in Figure 1 on Page 13.

In total, this process had the potential to produce 13,967,550 unique compounds, discounting those that were themselves present in the attested dataset. Given the sheer immensity of this pool, this process is assumed to be 'random enough'; in other words, the generated compounds should represent a substantial cross-section of the possible heads, modifiers, and semantic relationships of noun compounds as a linguistic structure.

The exact number of compounds (250) was chosen so as to match the size of the dataset used by Nakov and Hearst [27], and to fit within the time and financial constraints of this thesis.

Search-Engine Verification

As an additional step, the generated compounds were filtered using search-engine-based techniques to avoid pairings that may have been invalid for a number of reasons. These compounds may have involved irregular usage of acronyms or exceptional grammatical structure that could confuse human judges or otherwise compromise the analysis.

In particular, for each generated compound, a query was run on the Bing search engine to find instances in which the modifier and head appeared adjacently. The top 100 search

results were collected and shallow parsing was performed on their relevant snippets using NLTK's regular-expression-based parser [4]. Each compound was assigned a score that tallied the number of snippets (out of the top 100) for which the modifier and head appeared as unseparated nouns. For example, given the compound *baseball game*, the snippet "go to a baseball game..." was considered valid, while the snippet "I like baseball. Game..." was invalid.

Compounds with fewer than five valid snippets (out of the top 100 results) were discarded. As the bar for qualification was incredibly low, this filtration step was considered weak enough such that the average human judge would not have seen most compounds satisfying the criteria. In other words, while this filtration helped remove irregular noun compounds, it was *not* strict enough to bias the pool of candidate compounds towards those with which human judges would be familiar. This would be a valid concern had we required, say, 95 or more valid snippets (out of the top 100 results), as any noun compound that passed the test would by definition occur quite frequently. However, this low bar was chosen so as to preserve the randomness and unfamiliarity of the generated compounds.

The final list of 250 compounds generated by this process can be found in Section B.1 of the Appendix.

4.2 Human Intelligence Task Format

The Human Intelligence Task (HIT) template used in this first round of experiments was designed so as to collect two key pieces of information. Specifically, for each judgment, we required:

1. *An interpretability label.* This label had to be chosen based on the scale introduced in Section 2.4, allowing Turkers to choose from the *No difficulty*, *Minor difficulty*, and *Meaningless* designations. Note that Turkers were provided definitions nearly identical to those listed in Section 2.4.

2. *A paraphrase representing and explaining the Turker's interpretation of the noun compound.* A paraphrase was required if and only if the Turker deemed the compound to be interpretable with *No difficulty* or *Minor difficulty*, as *Meaningless* compounds by definition should not be paraphrasable.

There were several reasons for collecting paraphrases:

- *To prevent against low-quality work and keep Turkers honest*: By requiring a paraphrase, Turkers could not merely guess an interpretability label and pass it off as an honest attempt.

- *To force the Turkers to think deeply about the interpretability of the compound*: By going through the exercise of generating a valid paraphrase, Turkers were forced to consider whether the compound was legitimately interpretable, and to what degree.

- *To allow for synset annotation in the future*: With a paraphrase, the essence of the Turker's interpretation is captured, allowing for the compound's head and modifier to be assigned appropriate WordNet synsets, as described in Section 6.1.

- *To allow for paraphrase analysis in the future*, in consistency with the hypotheses outlined in Section 3.

When providing a paraphrase, Turkers were asked to use a fill-in-the-blank format. In particular, Turkers were given a template for paraphrasing the compound based on use of the relative clause or a preposition. For example, given the compound *pressure dispute*, Turkers were asked to fill in the following blank: "a pressure dispute is a dispute that [...] pressure(s)".[4]

This format is useful in that the resulting paraphrases are easy to analyze programmatically and easy for Turkers to produce. At the same time, it is sufficiently flexible as Turkers have access to an arsenal of verbs and prepositions. As seen in the work of Nakov and Hearst [27], use of this format with paraphrasing verbs is impressively powerful when it comes to noun compound analysis; similarly, Lauer [20] demonstrated that prepositions can be useful as well.

For an example of an HIT presented to Turkers, see Section B.2 in the Appendix.

4.3 Structure

For each of the 250 generated compounds constructed by the process defined in Section 4.1, three judgments were collected by Turkers, where each judgment conformed to the HIT format from Section 4.2. This made for 750 approved HITs in total.

In this section, we describe the manner in which these HITs were divided up among Turkers, and how their submissions were monitored and approved. On the AMT platform, quality control is essential [15]; as such, we designed our experiment so as to maximize the quality of submissions and avoid biasing our results towards the tendencies of any specific Turkers.

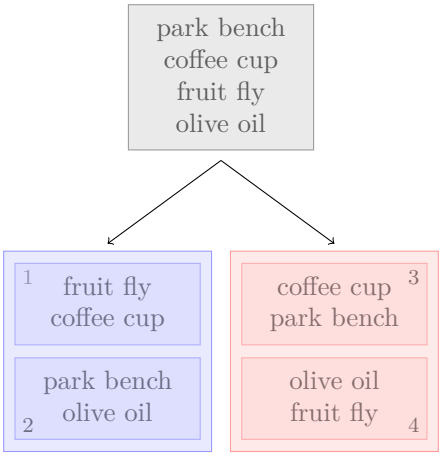
Batching

HITs were released in batches of 50 such that each of the 250 compounds appeared exactly once in the first five batches, exactly once in the second five batches, and exactly once in the final five batches.[5] In this way, each set of five batches constituted an individual run of the experiment, and the three sets of five batches constituted three such runs. The batching process is depicted in Figure 3.

[4] In the instructions, it is explained that 'that' can be replaced by a preposition, like 'who', 'where', 'of', and so forth.

[5] Due to an error, the first five batches were released as a single batch of 250 compounds. This should have little bearing on the experimental results, as the next two judgments were gathered in the correct manner. As such, each judgment collected was randomized relative to one another.

Figure 3: An example of the batching process, in which a set of four compounds is divided up into two sets of two batches. Note that each set of batches contains each of the four compounds exactly once. Batches 1 and 2 (blue) would be released first, in order, followed by Batches 3 and 4 (red). At completion, each compound would be judged exactly twice, and in a randomized order.



For each of these rounds of five batches (i.e., each set of 250 HITs), the order and grouping of compounds was randomized. In this way, no two judgments for a single compound occurred in the same 'surroundings', i.e., in the presence of the same group of compounds. This prevented against certain compounds being biased by their position in the set of HITs. In other words, it made for a random order of presentation, a crucial form of experimental control.

As an example of why this kind of batching can be helpful, consider a scheme that does not randomize presentation, and assume that we have one compound, A, that is very difficult to interpret and another compound, B, that is also relatively difficult to interpret, but slightly easier than the first. If we collected three judgments for these two compounds, each time presenting Turkers with A and then B, the Turkers could be biased by the difficulty of A and rate B as easier to interpret than it truly is. Thus, the concern is that a consistent order of presentation could affect the interpretability label assigned to a given compound. Through the batching process, we avoid these issues entirely.

Turker Restrictions

When conducting experiments on the AMT platform, one runs the risk of biasing results in favor of certain Turkers that submit a large portion of the available HITs, an issue that can be compounded by a lack of quality control. For example, in the context of this experiment, one runs the risk of having a single, very creative Turker submit a large proportion of the total judgments in which every compound is deemed interpretable, which may not be representative of the standard human judgment.

To prevent against such issues, we levied the following restrictions upon Turkers partici-

pating in our experiment:

- No Turker was allowed to submit multiple judgments for the same compound, thus guaranteeing that the interpretability of each compound was assessed by three distinct judges.
- No Turker was allowed to submit more than 50 judgments in total.
- Turkers could only participate if they had an approval rating above 99% and more than 500 HITs approved over their lifetime. (Note that experimenters can reject submissions from Turkers if they fail to follow instructions or otherwise represent low-quality work.)

These restrictions guaranteed a variety of judgments for a single compound and protected against any biases or tendencies present in the Turkers themselves.

Criteria for Rejection

The AMT platform provides the ability for experimenters to reject HITs submitted by Turkers if they fail to adhere to instructions or otherwise represent low-quality work. In this experiment, HITs were rejected if they:

- Provided a paraphrase after indicating that the compound was *Meaningless*.
- Interpreted either the head or modifier as a proper noun, adjective, or any other inappropriate part-of-speech.
- Misunderstood a word in the compound (e.g., interpreted 'statue' as 'statute').
- Misspelled a word in their paraphrase (an indicator of laziness).
- Provided a nonsensical paraphrase.

These criteria again demonstrate the value of collecting paraphrases, as low-quality work was easy to identify.

In some cases, Turkers who submitted suspicious results (e.g., an usual number of *Meaningless* judgments) had their judgments disqualified. The bar for judgment quality was kept very high, as each of the 750 individual judgments was reviewed and approved manually.

5 Overview of Results

In this section, we present the results of this first round of experiments from a high level. Namely, we focus on meta-data, such as the average length of time that a Turker spent on an HIT, and raw expository statistics computed over the submissions. In Section 6 below, we discuss the various ways in which the data was analyzed beyond these basic measures, such as through the use of advanced machine learning techniques, clustering algorithms, and dependency parsing.

5.1 Experiment Statistics

This first round of experiments was conducted in the window from October 19, 2014 to October 21, 2014.[6]

As we chose to use 250 distinct compounds and required three judgments per compound, we had to collect 750 individual HITs that satisfied the criteria defined in Section 4.3. In doing so, we rejected exactly 50 submissions, making for a 93.75% acceptance rate.

The average time spent by Turkers on an HIT was roughly 49.50 seconds. This number was well above our expected time per judgment (30 seconds). Note that Turkers are paid merely based on the number of HITs approved and not by the time spent on each task, which makes the 49.50 seconds a comforting result, as lengthier judgments would intuitively reflect more thoughtful consideration by Turkers.

Turker 'Diversity'

The 750 accepted HITs were submitted by a total of 83 different Turkers. The average number of accepted submissions per Turker was 9.04, while the median was 3, indicative of a long tail of Turkers that submitted just a few HITs (75% of Turkers submitted between 1 and 10 HITs). While it would have been advantageous to achieve a more even distribution, only 8 Turker submitted more than 30 HITs, and no Turker made it to the 50-HIT cutoff. In conclusion, the pool of judges was sufficiently diverse so as to avoid significant bias in the results, as per experimental design.

5.2 Breakdown of Submissions

Next, we introduce some basic, expository statistics about the results. In Section 6 below, we analyze the data in greater detail.

Of the 750 approved HITs, which spanned 250 distinct compounds, 299 submissions (39.9%) identified a compound as interpretable with *No difficulty*, 245 submissions (32.7%) identified a compound as interpretable with *Minor difficulty*, and 205 submissions (27.3%) identified a compound as *Meaningless*.

Given that we collected three judgments per compound, a logical next step was to view these three judgments as a vote on the compound's true interpretability label, and take the interpretability label receiving a majority of the votes to be the 'ground truth'. For example, if two judges deemed a compound to be interpretable with *Minor difficulty* and a third judge deemed it *Meaningless*, we would label the compound to be interpretable with *Minor difficulty*, given that two out of three judges agreed on that label. Note that if each of the three judges selected a different interpretability label for a given compound, this would result in a three-way tie and no clear majority; thus, the compound would be ineligible for this type of analysis. This occurred only in a minority of cases.

[6] In reality, 782 of the total 800 submitted HITs were completed in the aforementioned window. Due to revised rejection criteria, batches of size 5, 6, and 7 had to be re-run on October 23, November 5, and December 2, 2014, respectively.

By taking such a majority vote, we found that 96 of the 250 compounds (38.4%) were voted to be interpretable with *No difficulty*, 68 of the 250 (27.2%) were interpretable with *Minor difficulty*, and 55 of the 250 (22.0%) were *Meaningless*, leaving 31 compounds with no majority label.

Alternatively, by instead requiring *unanimity* among judges, we found that 43 of the 250 compounds (17.2%) were identified as interpretable with *No difficulty*, 18 of the 250 (7.20%) as interpretable with *Minor difficulty*, and 28 of the 250 (11.2%) as *Meaningless*.

These results are presented in Table 2.

Difficulty	**Num. Judgments**	**Num. Majority**	**Num. Unanimous**
No difficulty	299 (39.9%)	96 (38.4%)	43 (17.2%)
Minor difficulty	245 (32.7%)	68 (27.2%)	18 (7.20%)
Meaningless	205 (27.3%)	55 (22.0%)	28 (11.2%)

Table 2: Initial results from the Amazon Mechanical Turk experiments, which consisted of 750 approved HITs and 250 distinct noun compounds.

Interpretability as the Norm

Under the majority-vote criteria, nearly two-thirds of compounds (65.6%) were deemed interpretable with either *No difficulty* or *Minor difficulty*. If we restrict ourselves to those 219 compounds for which we had a clear majority-voted label, then 74.9% of compounds were deemed interpretable. This seems to validate hypothesis **H1** from Section 3, which claimed that the majority of compounds would be interpretable.

This high proportion is a testament to the productivity of noun compounds. As described in Section 4.1, the 250 compounds presented to Turkers were constructed randomly; most of these compounds were completely new to human judges. In general, one would expect such high entropy data to be riddled with noise. However, we instead found that these compounds were, more often than not, interpretable. The fact that human judges could ascribe meaning to almost three-quarters of these random compounds is a rather remarkable result.

Co-Agreement

In addition, we note that the co-agreement among judges was satisfyingly high for this experiment: 221 of the 250 compounds (88.4%) achieved majority agreement on an interpretability label, and 89 of the 250 compounds (35.6%) achieved consensus on an interpretability label.

Given the subjectivity inherent in the task, these values are seen as acceptable and representative of suitable experiment design. However, given that there has been very little academic research on the question of noun compound interpretability, we acknowledge that benchmarks are few and far between, making it difficult to contextualize the co-agreement figures presented above.

Majority vs. Unanimity

When determining a compound's interpretability label by majority vote, a larger proportion of compounds were identified as interpretable with *Minor difficulty* than as *Meaningless*. However, this result is reversed when requiring unanimity among judges.

When requiring the three judges to agree completely on a label, only 18 compounds were deemed interpretable with *Minor difficulty*, as opposed to the 43 compounds that were deemed interpretable with *No difficulty* (an increase by a factor of 2.39) and the 28 compounds that were deemed *Meaningless* (an increase by a factor of 1.56).

An alternative way of looking at this result: only 26% of majority-voted *Minor difficulty* compounds were agreed upon unanimously as being interpretable with *Minor difficulty*. Of that initial set of compounds, three out of every four had a dissenting judgment.

This is indicative of the notion that some compounds are obviously interpretable; others, obviously uninterpretable. But in between, there's a grey area where judges tend to disagree. Compounds deemed to be interpretable with *Minor difficulty* linger in this grey area, often characterized by the dissenting judgment mentioned above.

5.3 The Effect of Positioning on Interpretability

Recall that our AMT experiment was designed so as to minimize the impact that the ordering of HITs would have on the interpretability labels selected by Turkers. In that light, it is interesting to look at how submissions varied based on the positioning of an HIT within a batch. A related question, and one that we will look at in this section as well, is how submissions varied based on the number of HITs that a Turker had completed. For example, did Turkers become more open-minded as they saw more compounds, which could be indicated by a stronger preference for *No difficulty* and *Minor difficulty* judgments? Or was there some other pattern to the submissions?

When analyzing the results submitted by Turkers as a function of 'time', there are two possible lens that one can adopt. The first views time in terms of the position of an HIT within a batch; the second, in terms of the number of HITs that the specific human judge had completed before a given HIT. For example, given a batch with HITs `A`, `B`, and `C`, completed by judges J_1, J_2, and J_1 again, we could view `C` as the third HIT in a batch, or as the second HIT completed by its judge. These two lens (or perspectives) can also be phrased, respectively, as: plotting interpretability labels as a function of an HIT's position within a batch, and plotting labels as a function of Turker experience.

Given that HITs are presented to workers in a first-in-first-out order, one would expect these two approaches to capture similar ideas, as an HIT positioned later in a batch would more likely be completed by a Turker who had completed some of the earlier HITs. But the difficulty with the latter approach is that there was a large discrepancy in terms of the number of HITs submitted by Turkers, with many Turkers submitting just a handful of HITs, and a smaller subset submitting between 30 and 50 HITs. For completeness, we analyze the results from both lenses.

Across a Batch

When viewing time in terms of the proportion of a batch completed, there were no major discrepancies in labelling. In other words, the position of a compound in a batch did not seem to have a significant effect on the labelling of that compound.

Label	First Third	Second Third	Final Third
No difficulty	0.348	0.301	0.351
Minor difficulty	0.350	0.394	0.256
Meaningless	0.317	0.322	0.361

Table 3: The proportion of judgments for a given interpretability label that came in the first, second, and final third of HITs in a batch. For example, the top right cell indicates that 35.1% of *No difficulty* judgments came from HITs positioned in the final third of a batch.

Table 3 contains the proportion of judgments, for a given interpretability label, that occurred in a certain range of positions within a batch; in this case, batches are broken down into thirds. For example, 35.0% of *Minor difficulty* judgments came from HITs positioned in the first third of a batch (i.e., within the first 16 HITs, given that batches were of length 50). Given a perfectly uniform distribution, every cell in the table would read 0.33, indicative of a 33% split for each third of a batch.

The values in Table 3 roughly line up with expectation, which implies that position within a batch did not play a significant role in influencing interpretability labels. However, there are some minor deviations. For example, *Meaningless* judgments became more common near the tail end of a batch, and *Minor difficulty* judgments dropped from 39.4% in the second third to just 25.6% in the final third, suggesting that Turkers became less inclined to level *Minor difficulty* and more inclined to level *Meaningless* judgments as time went on. However, this does not seem to be a significant effect.

When considering these types of conclusions, it becomes clear that position within a batch is just a proxy for evaluating the effect that a Turker's experience played on the judgments they submitted. We analyze this effect in the next section.

Over a Turker's Lifetime

Next, we look at how the number of HITs completed by a given Turker influenced the interpretability labels of the judgments they submitted. Recall that in Section 5.1, we stated that 83 distinct Turkers had participated in our experiment, where 'participated' implies that they had at least one HIT submission accepted. Of those 83, 27 made exactly one acceptable judgment, while 56 made more than one acceptable judgment. In addition, note that Turkers submitted a total of 299 *No difficulty* judgments, 246 *Minor difficulty* judgments, and 205 *Meaningless* judgments, as presented in Table 2 of Section 5.

With that established, we present, in Figure 4, a plot of the number of judgments submitted for each interpretability label as a function of the experience of the Turker submitting the judgment; in particular, as a function of the number of HITs submitted by the Turker before

submitting the given HIT. To put it in simpler terms, each line in Figure 4 represents the rate at which judges with a certain level of experience (measured along the x-axis) deemed compounds to be interpretable with the labeling matched to that line.

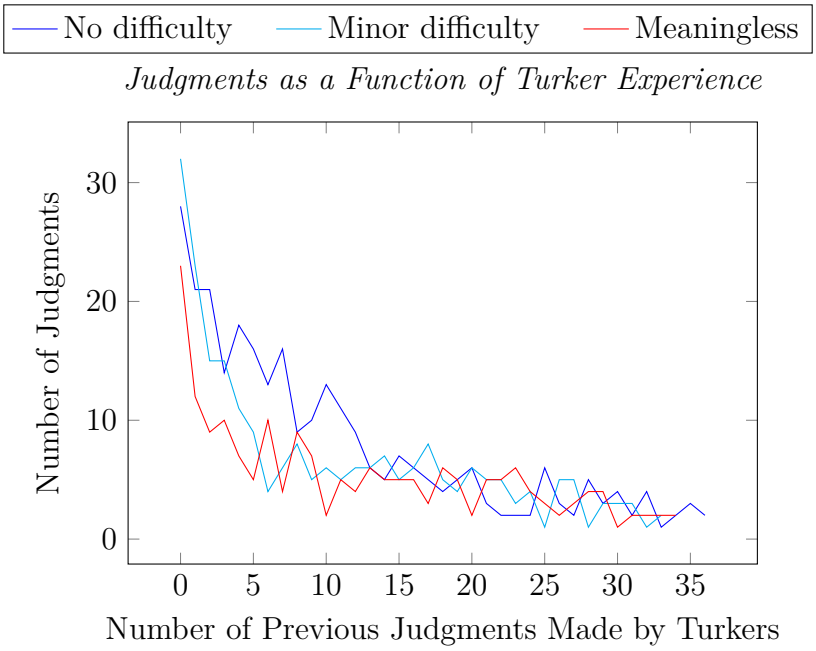


Figure 4: The number of judgments for a given interpretability label, as a function of the number of previous judgments submitted by the active Turker.

The figures presented in Figure 4 are *not* normalized, yet there are still more *Minor difficulty* judgments submitted as a Turker's first judgment than *No difficulty* judgments. To be precise, 32 Turkers submitted a *Minor difficulty* judgment as their first, while only 28 submitted a *No difficulty* judgment as their first, which respectively account for 13% and 9% of the total judgments submitted with those labels. The *No difficulty* judgments soon overtake the *Minor difficulty* judgments, but the initial discrepancy is curious.

We can couple this observation with the fact that the *Meaningless* judgment rates became increasingly competitive as Turkers submitted more and more HITs: after 10 or 15 HITs, the number of judgments for each interpretability label becomes roughly even, with the *Meaningless* label even containing more gross judgments for Turkers who had completed 18, 23, and 29 HITs at that point.

Figure 4, then, seems to depict a trend in which Turkers would initially shy towards *Minor difficulty* judgments and, over time, tend towards a higher rate of *Meaningless* judgments. This is likely connected to the idea of a *Minor difficulty* grey area, as described in Section 5: as it is likely that Turkers had not seen any of the generated noun compounds prior to their first judgment, they had no way to peg the difficulty of their first compound with respect to the other compounds in the dataset and thus opted for the 'safest' label, *Minor difficulty*, which sits between easy and hard. This again relates to the idea of *Minor difficulty*

compounds existing in a grey area between the obviously interpretable and the obviously uninterpretable.

While these trends are interesting, we feel that the randomization and other steps taken to ensure experimental hygiene (such as batching and rate-limiting for Turkers) were sufficient to yield high quality results. The discrepancies outlined in this section are relatively minor and, given that each judgment for a given noun compound was completed by a Turker with a different level of experience and placed in a different position within a batch, the interpretability labels determined by majority voting should still be valid for use as 'ground truths'.

6 Analysis

In the previous section, we presented some expository statistics from this first round of AMT experiments. Next, we apply a number of advanced techniques to analyze this data from various angles in an attempt to validate or invalidate the hypotheses from Section 3.

6.1 WordNet Sense Annotation

As mentioned in Section 4.2, one of the key motivations for collecting paraphrases from Turkers was to understand *how* they were interpreting noun compounds, a question that goes beyond the simpler task of identifying whether interpretation is possible.

This allowed us to compare the variety of interpretations attributable to a given compound and, in addition, compare those interpretations to attested noun compounds with similar structures. For example, while one Turker might paraphrase the compound *coffee cup* as "a cup that holds coffee", another might paraphrase it as "a trophy awarded for brewing the best coffee". These two paraphrases would represent completely different senses of the word *cup*: in the former, a *cup* is a container for holding liquid; in the latter, a trophy.

In Table 4, we present a sampling of the noun compounds used in our experiments. For each compound, we also include one of the paraphrases submitted by human judges, as well as the WordNet senses that best match the judge's interpretation, as determined by the paraphrase provided.

Compound	Paraphrase	Modifier Synset	Head Synset
machine actor	"An actor who works like a machine"	*machine.n.2*	*actor.n.1*
frog ring	"A ring that is decorated with frogs"	*frog.n.1*	*machine.n.8*
margin office	"An office that functions... at the margin"	*margin.n.1*	*office.n.1*
retirement practice	"A practice of a retirement ceremony"	*exercise.n.3*	*retirement.n.2*
string victim	"A victim that has been harmed by string"	*string.n.1*	*victim.n.1*

Table 4: Several generated noun compounds, along with a sample paraphrase and synset annotations.

As a more pertinent example, consider the generated compound *government eye*. In one of the judgments submitted by Turkers, we might find that the head, *eye*, was used in a manner similar to its function in the attested compound *private eye* (that is, to denote some sort of spy). In another, we might find that *eye* was interpreted as a physical human eye. Bent these two judgments, *eye* would have been used in very different senses and, as such, would need to be annotated with distinct WordNet synsets.

In order to make these comparisons, we chose to use WordNet senses as our unit of account. Specifically, for each generated noun compound, for each human judgment provided by Turkers, we used the paraphrase to assign a WordNet synset to the compound's head and modifier that best captured their respective usages.

WordNet Senses

In Section 2.3, we introduced the notion of a WordNet *synset*. WordNet synsets do not correspond to English-language words in a one-to-one manner: some English words might have multiple WordNet synsets, one for each *sense* in which the word might be used. As an example, the word *chair* maps directly to two synsets: *chair.n.1*, defined as "a seat for one person, with a support for the back", and *chair.n.5*, defined as "a particular seat in an orchestra".

Additionally, one can expand the set of candidate synsets to include those mapping to synonyms of a given word. These additional synsets would not be explicitly labeled with the given word, but would instead represent senses for which the word, when used in a certain way, would have the same semantic meaning. Returning to our *chair* example, if we query for all senses of the word *chair* including synonyms, then in addition to the synsets mentioned above, we get *professorship.n.1*, defined as "the position of professor" (as in, "Professor Appel is the chair of the Computer Science Department at Princeton.") and *electric_chair.n.1*, defined as "an instrument of execution by electrocution".

A Need for Manual Annotation

In academic research, it is common to default to WordNet's 'first sense' when determining an acceptable synset to represent a given word, especially in an automated manner. WordNet senses are ordered roughly by frequency of usage; thus, this heuristic can at times yield reasonable results [24].

However, the *government eye* example from the previous section demonstrates a need for manual sense annotation when working with noun compounds, where a subtle difference in usage can lead to a major difference in semantic meaning. For compounds, defaulting to the first sense provided by WordNet would lead to incredibly inaccurate synset annotation.

As such, we manually determined the optimal synset for the head and modifier of each compound, for each judgment collected on the AMT platform. As there were 750 total HITs, and each judgment required two synsets (one for the compound's head, another for its modifier), this required up to 1,500 assignments, making it a time-consuming but necessary task.

Sense Assignment

For the head and modifier, respectively, the list of candidate synsets was restricted to the set of noun synsets based on different senses of the head or modifier as well as its synonyms, determined through use of the Python NLTK library's WordNet interface [4].

The use of synonyms was deemed necessary in this case, given that the use of a word in a noun compound can vary greatly based on its position within the compound and the other words with which its paired. For example, if a Turker was presented with the generated compound *apple chair*, by allowing for the use of synonyms, we would be free to label the head (*chair*) with the synset *professorship.n.1* if they provided the paraphrase "An apple chair is the chair of a department that studies apples" or the synset *chair.n.1* if they provided the paraphrase "An apple chair is a chair that is made of applies".

Note that if a Turker judged a compound to be *Meaningless*, they were not required to submit a paraphrase; thus, *Meaningless* judgments could not be assigned WordNet senses, as there was no paraphrase from which to determine the optimal sense.

First-Sense Heuristic

In some cases, we did fall back to the 'first-sense' heuristic mentioned above. Namely:

- If a given word only had one WordNet sense available, including synonyms, we defaulted to that sense regardless of its appropriateness.
- If we ever needed to assign WordNet senses to a compound that was judged *Meaningless* and thus lacked a paraphrase, we defaulted to the first sense for its head and modifier.
- If we ever needed to assign WordNet senses to an attested compound, e.g., when comparing senses of generated compounds to senses of attested compounds.

This was a necessary procedure but was deemed acceptable for two reasons. First, in the scenarios listed above, perfect accuracy is more a luxury than a requirement. Second, in our annotation, we found that the first sense was used in a majority of judgments. To be precise: for heads in which there were multiple senses to pick from, we found that 56% of usages fit the first sense; for modifiers, the number was 66%. In other words, judges did end up using the first sense a majority of the time in their interpretations.

The results of this WordNet sense annotation process will be used at-length in the analysis that follows.

6.2 Diversity by Difficulty

In this section, we investigate the hypotheses of Section 3 that relate to the diversity of interpretations for a given compound and, in particular, analyze this diversity as it relates to the difficulty of interpretation, as indicated by Turkers.

As a reminder, these hypotheses primarily claimed that:

- Compounds that are easier to interpret should tend towards fewer possible or 'best' interpretations (**H2**).
- Compounds that are easier to interpret should require shorter paraphrases with less diverse tokens, e.g., more prepositions and fewer verbs, given the expressive power of verbs (**H5**).

These hypotheses are rooted in the principle of Occam's Razor, as they assume that the human brain favors simpler interpretations, when possible. Further, these hypotheses, if validated, would suggest that ambiguity is a complicating factor when interpreting new compounds.

As we were unable to collect paraphrases for *Meaningless* compounds, by virtue of their definition, this section focuses primarily on the differences between compounds judged to be interpretable with *No difficulty* or *Minor difficulty*.

Diversity of Interpretations

To begin, we look at the diversity of interpretations submitted for a given compound, as judged by the number of different synsets assigned to the compound's head and modifier. We divide our compounds into those for which a majority of judges deemed them to be interpretable with *No difficulty* and those for which a majority deemed them to be interpretable with *Minor difficulty*. As per Section 5.2, 96 and 68 compounds fit these definitions, respectively. The remaining 86 compounds were ignored.

We computed the average number of senses per compound for the *No difficulty* and *Minor difficulty* compounds, both for their heads and modifiers (recall that each compound had between two and three paraphrases, as a majority of the judgments had to be of either *No difficulty* or *Minor difficulty*, leaving room for at most one *Meaningless* judgment with no sense annotation). Additionally, we computed the percentage of compounds for which each interpretation used the same sense–in other words, those compounds on which human judges provided a uniform, universally-agreeable interpretation. The results are presented in Table 5 below.

	Difficulty	**Mean Senses**	**Single Sense**
Head	None	1.146	86.5%
	Minor	1.324	72.1%
Modifier	None	1.135	86.5%
	Minor	1.191	80.9%

Table 5: The average number of senses in which the heads and modifiers of a set of compounds were interpreted, segmented by difficulty of interpretation. The percentage of compounds for which the head or modifier were used in a single sense is listed as well. Compounds judged to be interpretable with *Minor difficulty* showed a more diverse range of interpretations, across both heads and modifiers, than those judged to be interpretable with *No difficulty*.

As shown in Table 5, *Minor difficulty* compounds demonstrated, on average, greater diversity in interpretation, as reflected by increases of 15.5% and 4.9% in the mean number of WordNet senses for heads and modifiers, respectively. Similarly, human judges were much more likely to gravitate towards a universal interpretation for *No difficulty* compounds than for *Minor difficulty* compounds, as seen in the increased single-sense percentage for *No difficulty* compounds.

The results in Table 5 seem to validate hypothesis **H2** from Section 3 in that compounds that were more difficult to interpret yielded a wider variety of interpretations and greater ambiguity; whether this is a causal relationship remains unclear.

It is interesting to note that diversity was maximized for the heads of the *Minor difficulty* compounds. While this was an unforeseen result, we view it as further evidence of a relationship between diversity and difficulty. As the name suggests, the 'head' of a noun compound is the anchoring noun: it provides the base of the entity described by the series of words. For example, the compound *cider apple* describes a type of apple, and the compound *jungle village* describes a type of village. Thus, diversity of heads should be more indicative of semantic diversity than diversity of modifiers, and this intuition is reflected in Table 5.

Diversity of Paraphrases

Next, we examine diversity with regards to the paraphrases submitted by Turkers.

As described in Section 4.2, each HIT contained explicit instructions as to how Turkers should structure their paraphrases. Turkers were given strict instructions: their paraphrases had to revolve around a relative clause or preposition. For example, a Turker presented with the compound *abbey assembly* would be asked to fill in the phrase: "an abbey assembly is an assembly that/of/from [...] abbey." Submissions for this particular compound included: "gathers at an", as in, "an abbey assembly is an assembly *that gathers at an* abbey"; and "of the people who work or live at an", as in, "an abbey assembly is an assembly *of the people who work or live at an* abbey".

This format made it easy to analyze paraphrases programatically. In order to standardize submissions, we wrote a short script to clean the submitted paraphrases and coerce them into the format described above, taking into account, for example, that some Turkers included the surrounding copy in their submissions, while others just filled in the blank (e.g., some Turkers might submit "an abbey assembly is an assembly that gathers at an abbey", while others would simply submit, "gathers at an"). Others still deviated in non-substantial ways from the required format, e.g., by preceding their paraphrases with fragments like "might be" or "is probably"; these deviations were also smoothed out by the cleaning script.

As in the previous section, we grouped compounds by taking a majority vote over the interpretability labels provided by Turkers. To enrich the analysis, we also introduced a new category of compounds: *Eccentric*. This category included any compound that exactly two human judges deemed *Meaningless*, with the third judge deeming it interpretable with *No difficulty* or *Minor difficulty*.[7] Given that two out of three human judges were unable to

[7]Note that *Eccentric* compounds would also be categorized as *Meaningless* compounds when using a

interpret these compounds, the third judgment is of particular interest, as it likely required significant creativity on the part of the interpreter, and this creativity should be embodied by the relevant paraphrase.

In a way, *Eccentric* compounds are as close as we can get to including *Meaningless* compounds in our paraphrase-based analysis, making them especially useful. However, only 27 compounds qualified as *Eccentric*, so their use is purposefully limited in our analysis.

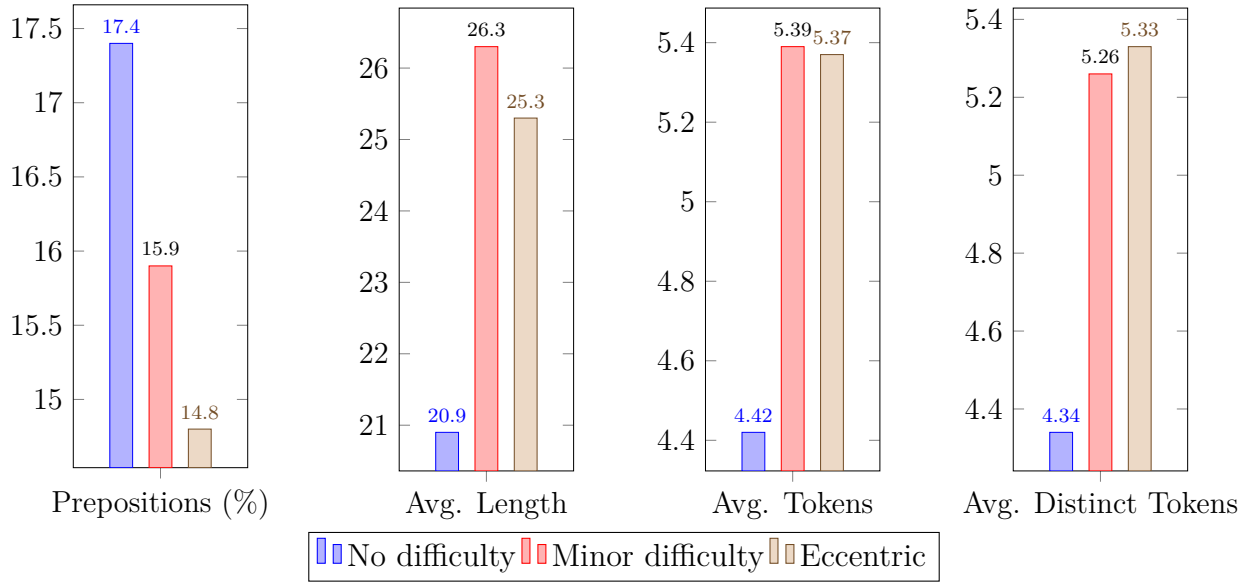


Figure 5: Diversity of paraphrases, as judged by four different metrics and assessed over compounds as grouped by difficulty of interpretation. As compounds become more difficult to interpret, human judges composed more advanced paraphrases, using fewer prepositions and a greater number of tokens.

Recall that paraphrases submitted by Turkers took the form "`modifier head` is a `head` that [...] `modifier`", where the word 'that' could optionally be replaced by a preposition. When analyzing paraphrases, we removed the surrounding context to extract the content in the blank, as well as the leading preposition or the word 'that'. This measure was taken to remove biasing our results towards longer compounds. For example, if we merely looked at the length of paraphrases, then those paraphrases involving the compound *hydrogen refrigerator* would of course be longer, on average, than those involving *ant hill*, by virtue of including a longer compound. In the analysis that follows, then, when we refer to 'paraphrases', we're referencing the content used by Turkers to fill in the blank, as well as the leading preposition or 'that'.

In analyzing the diversity of paraphrases, we computed four key metrics:

- *Average length*: A measure of the raw number of characters in a paraphrase.

majority vote.

- *Average number of tokens*: A measure of the raw number of words in a paraphrase.
- *Average number of distinct tokens*: A measure of the number of unique words in a paraphrase.
- *Percentage of paraphrases using prepositions*: As prepositions come from a fixed set (unlike verbs, which are theoretically unbounded in number), they are seen as a simpler method of paraphrasing. As such, increased use of prepositions is indicative of compounds that are easier to paraphrase and thus easier to interpret.

The values of these four key metrics, computed across the *No difficulty*, *Minor difficulty*, and *Eccentric* compounds can be found in Figure 5. As is clear from the graphs, paraphrase complexity and diversity correlated with difficulty of interpretation, as measured every metrics considered above. This correlation is consistent from *No difficulty* to *Minor difficulty* compounds and from *Minor difficulty* to *Eccentric* compounds, with the exception of a very minor length increase from *Eccentric* to *Minor difficulty* compounds.

The use of prepositions, for example, decreased as difficulty of interpretation increased. From the submissions, it is clear that preposition usage is indicative of simplicity of interpretation, especially for *No difficulty* compounds. For example, the compound *mother requests* was deemed to be interpretable with *No difficulty*, and was paraphrased by one Turker as: "mother requests are requests of a mother". The paraphrase is simple and clear, reflective of the compound itself, which is relatively straightforward. Similarly, the compound *hermit committee* was judged to be interpretable with *No difficulty* and paraphrased as: "a hermit committee is a committee of hermits". Again, the simple paraphrase is indicative of the ease with which the compound was interpreted.

On the other hand, take one of the *Eccentric* compounds, *siphon letter*, paraphrased by one Turker as: "a siphon letter is a letter that draws attention away from other letters in a word". The core content of this paraphrase includes nine distinct tokens and, indeed, the need for such an elaborate explanation is evident given the complexity of the compound.

The results in Figure 5 validate hypothesis **H5** from Section 3, which claimed that compounds that are more difficult to interpret would demand more complex paraphrases. Intuitively, this finding is in sync with the manner in which we provide explanations in the real world: concepts or compounds that are difficult to understand are also difficult to explain, and, as such, require longer and more precise explanations.

6.3 Comparisons to Attested Compounds

The next step in our analysis was to look outside of the raw data collected on the AMT platform by expanding our scope to include comparisons to attested compounds. As defined in Section 2.4, attested compounds are those present in existing noun compound datasets. Given the manner in which these datasets were constructed, attested compounds are assumed to be easy to interpret and would often be familiar to human judges. For example, *bike company*, *data transfer*, and *fishing vessel* are all attested compounds.

Our set of attested compounds was composed of the union of the Kim and Baldwin [17], Nakov and Hearst [27], Ó Séaghdha and Copestake [31], and Tratz and Hovy [41] datasets, which range in size from over 18,000 to just 250 compounds. In total, this set included 20,710 distinct compounds.

Intuition: Why Compare to Attested Compounds?

To explain why attested compounds would be of interest to us in developing a theory of interpretability, we return to the Principle of Compositionality. As defined in Section 2.1, this Principle states that the meaning of a compound is a function of the meanings of its constituent components and the way in which they are syntactically combined. In the context of noun compounds, then, this Principle suggests that when interpreting a new compound, a human judge would first parse its modifier and head independently, and then find some way to combine their meanings.

Consider, then, encountering a new compound, like *cotton cup*. If we assume the Principle of Compositionality, a human judge might first parse *cotton* and then *cup*. In their head, they may think back to similar compounds following the pattern (∗ *cup*). For example, our set of attested compounds includes the compound *paper cup*. One might recall that *paper cup* describes a "cup made of paper" and, by noting that *cotton* and *paper* are semantically similar in that they both refer to materials, infer that a *cotton cup* could be a "cup made of cotton."

This intuition can be generalized even further to suggest that compounds of the form * *cup*, where the modifier is replaced by a material like *cotton* or *stone*, are easier to interpret than if we'd filled in the modifier with some other random word. In the end, this is a function of the semantic similarity and the shared category membership of the modifiers in *cotton cup* and *paper cup*.

In the context of our experiments, then, the question becomes: *Can we model difficulty of interpretation as a function of the semantic similarity between generated and attested compounds?* As in the example above, is *cotton cup* easier to interpret than, say, *jungle cup*, given that, out of *cotton* and *jungle*, the former is semantically closer to *paper*. More generally, given a generated compound, can we draw inferences about its difficulty of interpretation by comparing it to the attested compounds that share either the generated compound's head or a modifier?

Evidence in support of such a model would be, by extension, evidence in support of an approach to noun compound interpretation that makes use of familiar compounds and attempts to draw links between the old and the new. In effect, such an approach could be viewed as an *exemplar-* or *nearest-neighbor-based* model of interpretation: when given new compounds, we search for semantically similar compounds with which we're already familiar and rely on those when developing novel interpretations.

Methodology

Drawing on the intuition from the previous section, we modeled semantic similarity between generated and attested compounds using head- and modifier-based comparisons based, separately.

For example, when comparing based on the similarity of modifiers, we would compare *cotton cup* to *paper cup* as, in this case, we're concerned with the semantic similarly of the modifiers (*cotton* and *paper*) over compounds that share a head (*cup*).

Alternatively, when comparing based on the similarity of heads, we compare *cotton cup* to, say, *cotton shirt* or *cotton farmer*, as we're concerned with the similarity of the heads (*cup* and *shirt*) over compounds that share a modifier (*cotton*).

Compound	Modifier Variant	Head Variant
cotton cup	coffee cup	cotton farmer
lightning country	wine country	lightning bolt
bear helmet	steel helmet	bear bone
dinner officer	prison officer	dinner guest
beard alcohol	grain alcohol	beard trim

Table 6: Modifier and head variants for a set of generated compounds, where modifier variants require a shared head, and head variants require a shared modifier. In both cases, variants must be drawn from the set of attested compounds, i.e., those with which human judges would typically be familiar.

We refer to these constructs as **modifier variants** and **head variants**, respectively, such that *paper cup* is a modifier variant of *cotton cup* and *cotton farmer* is a head variant of *cotton cup*. Several additional examples are presented Table 6. Note that variants must be drawn from the pool of attested compounds. To simplify the experiment, we chose to analyze the effects of modifier and head variants separately.

The synsets used for the generated and attested compounds were produced through the process described in Section 6.1.

As discussed in Section 2.3, semantic similarity-based comparisons often rely on similarity metrics built on top of WordNet. Recall that WordNet is modeled as a tree of *synsets*, which represent semantic concepts. WordNet-based semantic similarity metrics typically look at the distance between nodes in the WordNet graph using different definitions of 'distance'. For example, the simplest metric, *shortest-path distance*, simply counts the length of the shortest path between two synsets (nodes), which makes for a reasonable measure of similarity. For example, the synsets *paper.n.1* and *cotton.n.1*, defined as "a material made of cellulose pulp derived mainly from wood or rags or certain grasses" and "soft silky fibers from cotton plants in their raw state", respectively, have a shortest-path distance of just 5. On the other hand, the synset *jungle.n.3*, defined as "an impenetrable equatorial forest", is a distance of 13 away from *paper.n.1*. These distances are reflective of the semantic similarity between *paper* and *cotton*, at least in the material senses of the two words, and the lack thereof between *paper* and *jungle*.

In addition to *shortest-path distance*, we made use of some more complicated metrics, like *Wu-Palmer Similarity*, which is based on the depth of the two senses in the WordNet taxonomy and the depth of their most specific ancestor node. The full list of evaluated metrics can be found in Table 7 [25].

Name	Acronym	Description
`Shortest-Path Distance`	SP	Length of shortest path
`Shortest-Path Similarity`	PS	Maximum depth minus length of shortest path
`Leakcock-Chodorow`	LCH	Length of path, accounting for maximum depth
`Wu-Palmer`	WP	Length of path, accounting for common ancestor
`Resnik`	RES	Information content similarity
`Lin`	LIN	Semantic distance based on information content

Table 7: WordNet-based semantic similarity metrics and their respective definitions.

Overview of Variants

In comparing our generated compounds to attested compounds with which human judges would often be familiar, we developed the concepts of head and modifier variants, which rely on comparison between compounds that share either a modifier or a head, respectively. As a given compound can have multiple head and modifier variants, we briefly discuss some metrics related to variants of *No difficulty*, *Minor difficulty*, *Eccentric* compounds, with figures presented in Table 8.

Compound Type	Total Compounds	Avg. Head Variants	Avg. Modifier Variants
No difficulty	96	31.896	24.104
Minor difficulty	68	15.985	23.132
Eccentric	27	11.111	14.889

Table 8: The average number of head and modifier variants for *Eccentric* compounds, as well as those labeled interpretable with *No difficulty* and *Minor difficulty*, based on a majority vote over the judgments submitted on the AMT platform.

As seen in Table 8 above, the *No difficulty* compounds had, on average, over 31 and 24 head and modifier variants, respectively, while the *Eccentric* compounds had just 11 and 14. In other words, the average *No difficulty* compound shared a modifier with 31 attested compounds, and a head with 24, while the average *Eccentric* compound shared a modifier with just 11 and a head with just 14 attested compounds.

The compound with the most head variants was *government power* (as well as the three other generated compounds with the *government* modifier), with 246 attested compounds sharing that modifier. The compound with the most modifier variants was *top group* at 211 variants. Every compound had at least one head and at least one modifier variant, although 26 and 39 compounds had *exactly* one head and modifier variant, respectively.

That the number of variants correlates inversely with difficulty is an interesting result, and one that is in sync with the notion that human judges look for familiar variants when interpreting new compounds: if a compound had a greater number of variants, one would expect that it would be easier to find an attested compound semantically similar to a generated compound. In other words, if a head, for example, is used more commonly in noun compounds in the wild, it likely fits into a greater variety of semantic relationships, and thus a generated compound based on that head is more likely to be interpretable. In this way, the number of variants can be viewed as a measure of the flexibility of the fixed component.

One might be tempted to view the number of variants for a given word as a proxy for its polysemy, or the number of WordNet synsets in which the word occurs and thus the number of senses in which it can be used. And, indeed, when constructing modifier variants, the number of variants does correlate positively with the number of senses in which the fixed head occurs. In particular, these two quantities exhibit a Pearson correlation coefficient of $r = 0.1491$ and a two-tailed p-value of 0.0274, therefore withstanding a 5% significance test. Yet this correlation does not hold when comparing the number of head variants to the polysemy of the fixed modifier, as the p-value for these two quantities reaches 0.6670. Thus, while the number of variants is likely linked to polysemy, the concepts are distinct, with the former more a measure of the frequency at which a given word occurs within the general pool of noun compounds and the diversity of the semantic relationships in which it can participate, and the latter, a measure of the raw senses in which it can be interpreted (i.e., in isolation, outside of the context of noun compounds).

Modeling Difficulty as a Function of Distance

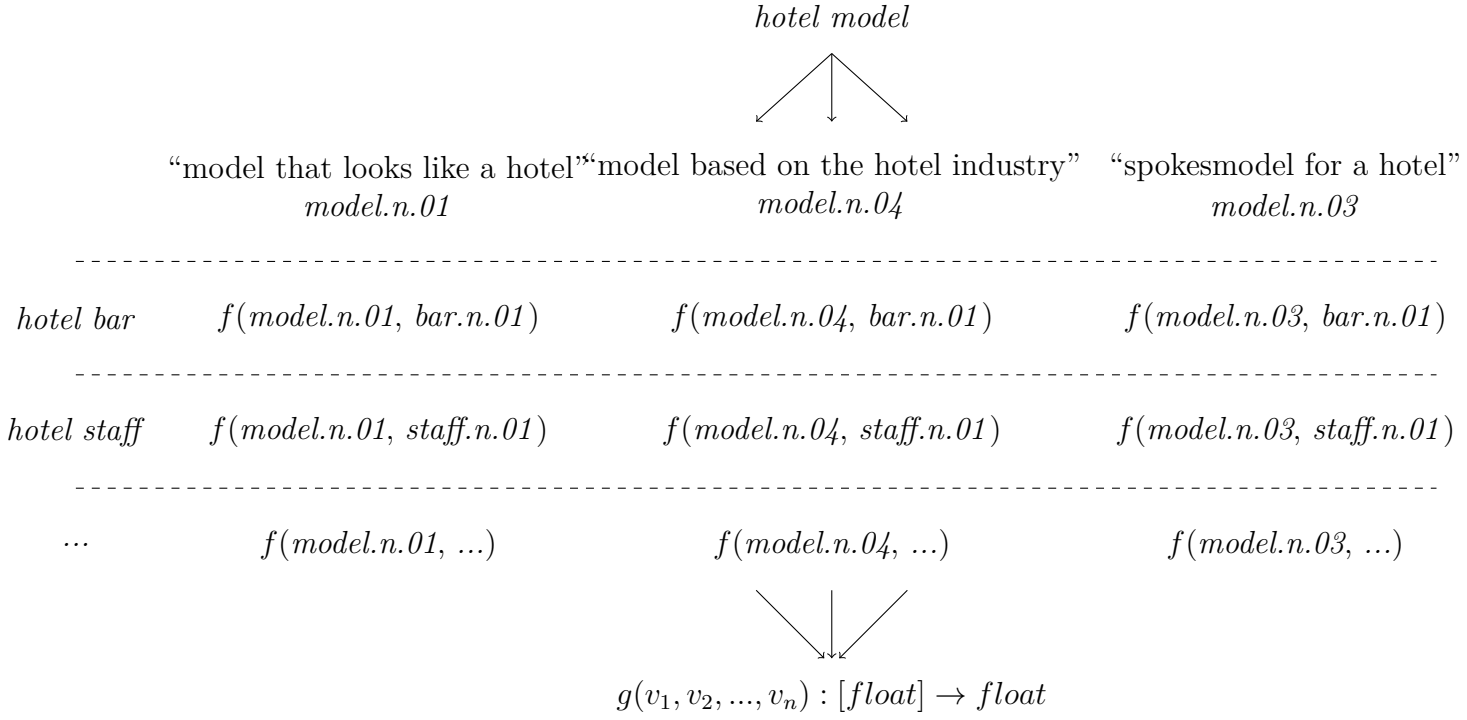
Next, we apply our WordNet-based semantic similarity metrics to the analysis of generated noun compounds by relating them to their head and modifier variants. The algorithm for computing similarity proceeded as follows: for each compound, for each judgment,[8] for each attested variant (either head or modifier variants, depending on the experimental configuration), we compute a semantic similarity vector between the synset assigned to that judgment and the synset of the attested variant. The similarity vectors were then averaged to produce a raw similarity score for the given compound type, be it *No difficulty*, *Minor difficulty*, or *Eccentric*. The complete process is demonstrated in Figure 6 using the generated compound *hotel model* and real paraphrases collected on the AMT platform.

Note that each compound will be represented by multiple similarity vectors, since we're computing a similarity vector for every pair of judgments and attested variants. A compound could have as many as three different judgments (i.e., three different synsets corresponding to the three paraphrases provided by Turkers) and hundreds of attested variants, making for hundreds of similarity vectors per compound, in some cases.

As such, we had to take care in computing a raw score; depending on our methodology, combining these vectors in different ways could yield different results with different

[8]Note that if two judgments for a given compound end up using the same WordNet synset in their interpretations, we include their similarity vectors twice to properly weigh the significance and relevance of each interpretation.

Figure 6: The process through which similarity vectors are generated, where f is a function that takes two WordNet synsets and produces a similarity vector, and g is a function that takes a set of similarity vectors and combines them to compute a raw score. Here, the generated compound *hotel model* is interpreted in three different ways by Turkers: first, as a physical, miniature model; second, as an abstract model; and third, as a physical, human model. Each of these senses of the word *model* is then compared to the heads of various attested compounds that share the modifier *hotel*, such as *hotel bar* and *hotel staff*.



implications. We examine three such approaches to vector combination:

- *Cumulative average*: In this model, we treated each similarity vector as an independent data point. In other words, we simply computed the average for each metric over every vector with no concern for which vectors corresponded to which compounds. This is the most straightforward approach, but is susceptible to bias in that compounds with more attested variants have a greater bearing on the final score. For example, while one compound might produce a hundred data points, another could produce as few as three; yet as each data point was treated independently, the first compound would have a much more significant influence on the final score.

- *Average of averages*: In this model, we first averaged all of the vectors for a single compound and then computed an average over the vectors of averages. This ensured that each compound was weighed equally in computing the final score.

- *Average of best-vectors*: In some sense, averaging is an unfair process given the intuition

behind this analysis. In motivating the use of semantic similarity, we proposed that a human judge might seek out the 'best-fit' variant, i.e., that with which they're familiar, like the use of *paper cup* when interpreting *cotton cup*. Thus, it might be the case that the closest variant is far more important than the *average* variant. As such, in this model, we computed the score for each compound by taking its highest-similarity vector, where 'highest-similarity' is selected by computing the vector with the largest norm.[9] These most-similar vectors were then averaged.

For each of these three approaches, the results, for head and modifier variants respectively, are presented in Figures 7 and 8.

Analysis of Results

We now discuss the results presented in Figures 7 and 8 on Pages 39 and 40, respectively.

For each of the three approaches to computing an aggregate score, over every metric, and when using both head and modifier variants, the outcome is nearly universal, barring a few exceptions: *compounds that are easier to interpret are more semantically similar to their attested variants.*

The differences are most pronounced when using the *average of best-vectors* approach, where the average shortest path for modifier variants differs by as much as 1.628 between *No difficulty* and *Eccentric* compounds. This is consistent with the intuition that human judges search for the 'best' familiar compound when encountering a new, generated compound.

In some cases, there appears to be a large drop-off between *Minor difficulty* and *Eccentric* compounds, but only a small drop-off from *No difficulty* to *Minor difficulty* compounds. As in previous analysis, this is suggestive of the notion that *Minor difficulty* compounds float in a grey area, but *Meaningless* (or, in this case, *Eccentric*, our proxy for *Meaningless*) compounds are relatively clear-cut and are much more removed from the *No difficulty* compounds with which human judges are typically familiar.

Between usage of the head and modifier variants, results were reasonably consistent. In other words, neither the heads nor modifiers appeared to be significantly more helpful in modeling difficulty as a function of distance to attested variants. For head variants, the gaps between *No difficulty* and *Minor difficulty* compounds across the similarity metrics do appear to be slightly larger, but the trends (decreasing similarity as difficulty increases) are slightly more clear-cut when using modifier variants.

In general, the results seen in Figures 7 and 8 seem to validate several of the hypotheses and ideas discussed above, namely **H3** from Section 3, which claimed that semantic similarity metrics and comparisons to attested compounds would be useful in gauging the interpretability of a compound. Indeed, semantic similarity does seem to be a useful indicator in assessing interpretability given the correlation between similarity scores and interpretability labels; in particular, the similarity of generated and attested variants goes a long way towards illustrating which compounds can and cannot be parsed by human judges. The usefulness of

[9] As the shortest-path distance metric operates such that a lower value indicates greater similarity, we took the inverse of the distance when computing the vector norm.

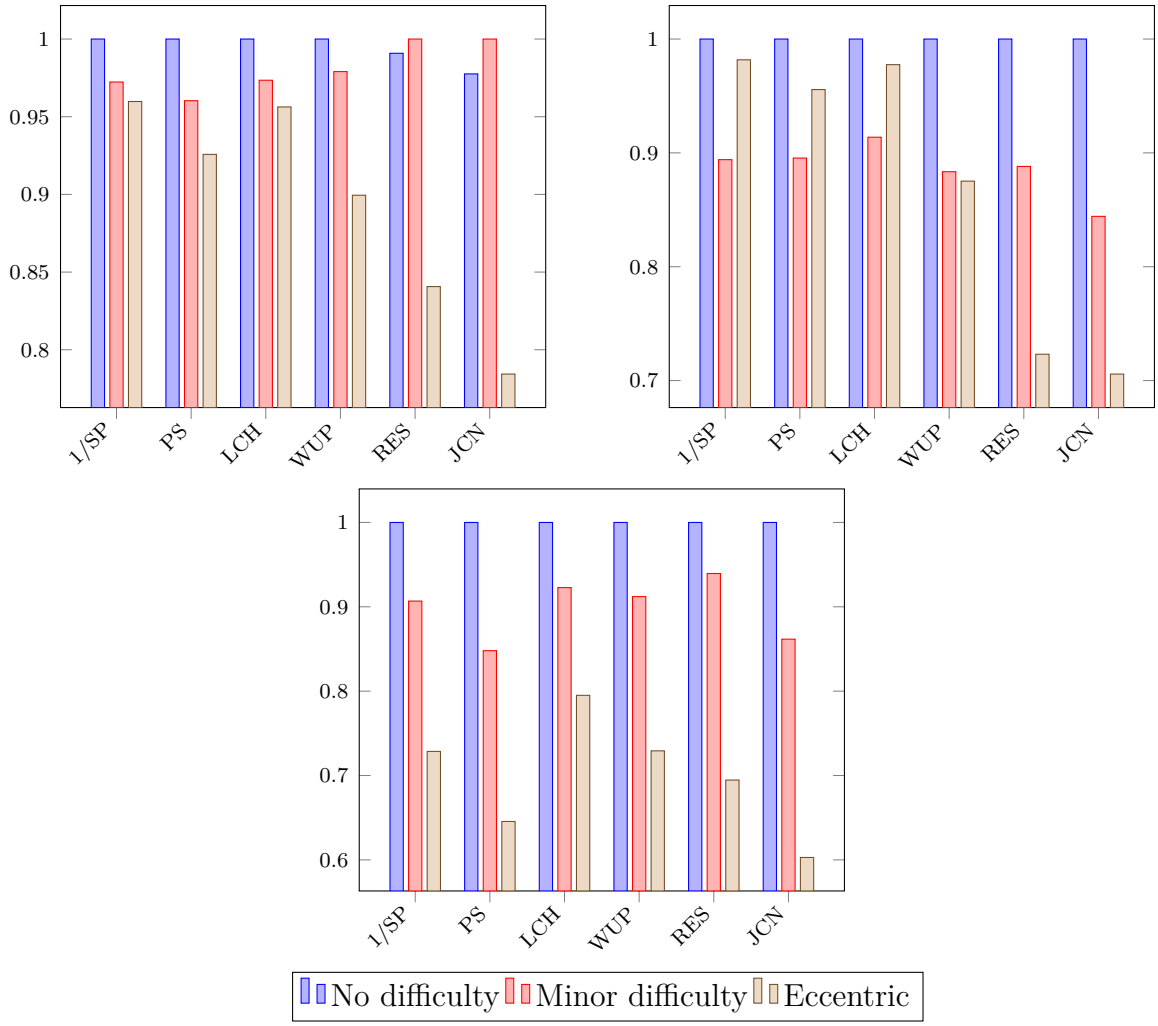


Figure 7: Results for the distance-by-difficulty analysis performed by comparing generated compounds to attested compounds with a shared *modifier* (also known as *head variants*). The techniques for each plot, clockwise from left, are *cumulative average*, *average of averages*, and *average of best-vectors*. In each plot, for each metric, scores are divided by the maximum value observed for that metric. We plot the reciprocal of the shortest-path distance so that, for each metric, larger values are indicative of greater similarity.

these semantic similarity features and comparisons to attested compounds will be evaluated further in Section 6.6 below.

But the results of Figures 7 and 8 are most interesting insofar as they inform us of the mechanisms by which human judges interpret new compounds. Based on the graphs below, it would appear that comparisons to existing, known compounds are key to parsing unfamiliar compounds and, as such, compounds for which such comparisons cannot be made are more difficult to understand.

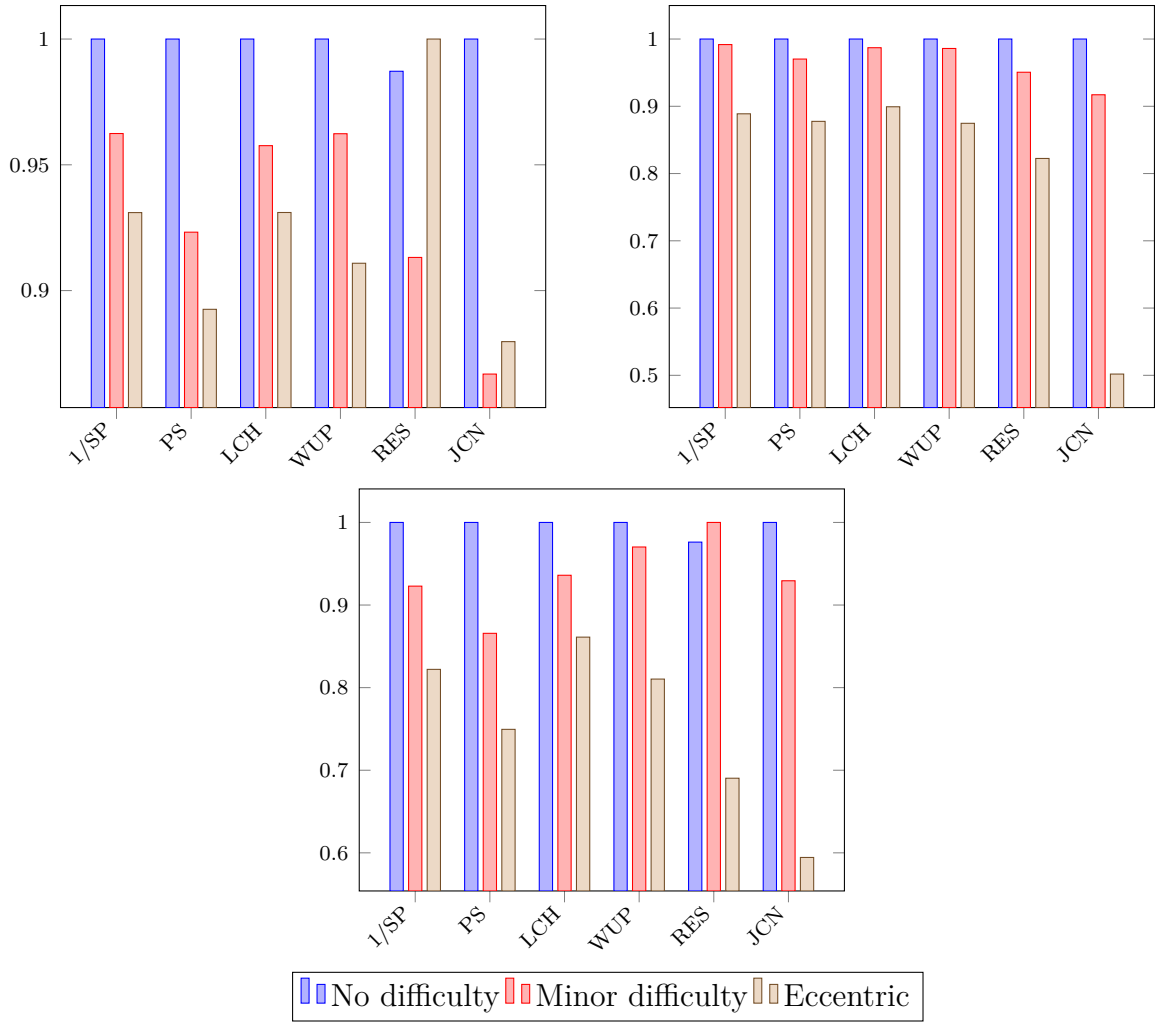


Figure 8: Results for the distance-by-difficulty analysis performed by comparing generated compounds to attested compounds with a shared *head* (also known as *modifier variants*). The techniques for each plot, clockwise from left, are *cumulative average*, *average of averages*, and *average of best-vectors*.

6.4 Modeling Difficulty as a Function of Word Frequency

An alternative approach to understanding a compound's interpretability label is to look at the frequencies at which its constituent components (i.e., its head and modifier) occur in corpora of English text. To put it simply, it could be the case that human judges consider words that they see infrequently to be more difficult to interpret and thus deem the compound (of which those words are a part) difficult to interpret as well. In some cases, the judge may not understand the word, given that it is used sparingly in English; in others, the judge could view the word as awkward or uncomfortable when used in a noun compound, or have a general lack of familiarity with how that word is used in context, leaving them unaware of

other compounds based on that word and incapable of drawing comparisons, the importance of which were demonstrated in the previous section.

In a way, this is an unsatisfying definition of difficulty. When we talk about difficulty of interpretation for noun compounds, we are trying to get at the semantic relationships between the various words that compose it. For a compound to be deemed uninterpretable by virtue of the relative infrequency at which its constituent components are used is less satisfying than, say, deeming it uninterpretable due to the inability to merge the respective meanings of its constituent components in a sensible manner.

At the same time, it would be unfair and unwise to underestimate the effect that word frequency can play on language understanding and production–an effect that has been demonstrated many times in the past, as evidenced by the work of Ellis [9]. Thus, in the context of this study, it is nonetheless important to look at the effect that frequency played on the interpretability labels assigned by human judges. This can be useful both to shed light on how judges made their decisions and to give us confidence in the significance of our results in later sections, when we make claims as to the role that semantic similarity and other factors played in influencing interpretability.

To start, we collected every word used as either a head or a modifier in any of the 250 generated compounds used in this first round of experiments. For each word in that set, we determined the frequency at which it appears in English text using the Google Ngram Viewer from Michel et al. [26]. Specifically, we treated each word as a unigram input, and the Google Ngram Viewer reported the percentage of text comprised of that word, as computed across a corpus of English books. The Google Ngram Viewer always returns frequency statistics for several different years; the value corresponding to the most recent year was preferred in each case.

The summary statistics for frequencies of the heads and modifiers of these 250 compounds, grouped by majority-voted interpretability label of the compound, are presented as box plots in Figure 9.

Based on the plots in Figure 9, it is immediately evident that frequency counts correlated with compound interpretability, to some degree. For each of the three interpretability labels, there was a large right skew in frequency counts for the words composing compounds with that label, which is consistent with the established observation that word frequencies follow a heavy-tailed distribution [40]. In other words, there was a large proportion of words with very low frequency counts, and a small proportion of words with significantly higher counts.

Although the correlation between frequency and interpretability is nontrivial, a substantial number of compounds deviate from the trend. For example, for both heads and modifiers of *Meaningless* compounds, there are four outliers (i.e., words with frequencies beyond $Q3 + 1.5 * IQR$) out of just 55 majority-voted *Meaningless* compounds, making for an outlier rate of over 7%. In fact, for heads, five different *Meaningless* word frequency counts would be above the third quartile for the *No difficulty* box plot (9.1% of the data points); and for modifiers, six different *Meaningless* data points would be above the *No difficulty* box plot's third quartile.

On the low end of the frequency spectrum, we see that the medians and lower quartiles

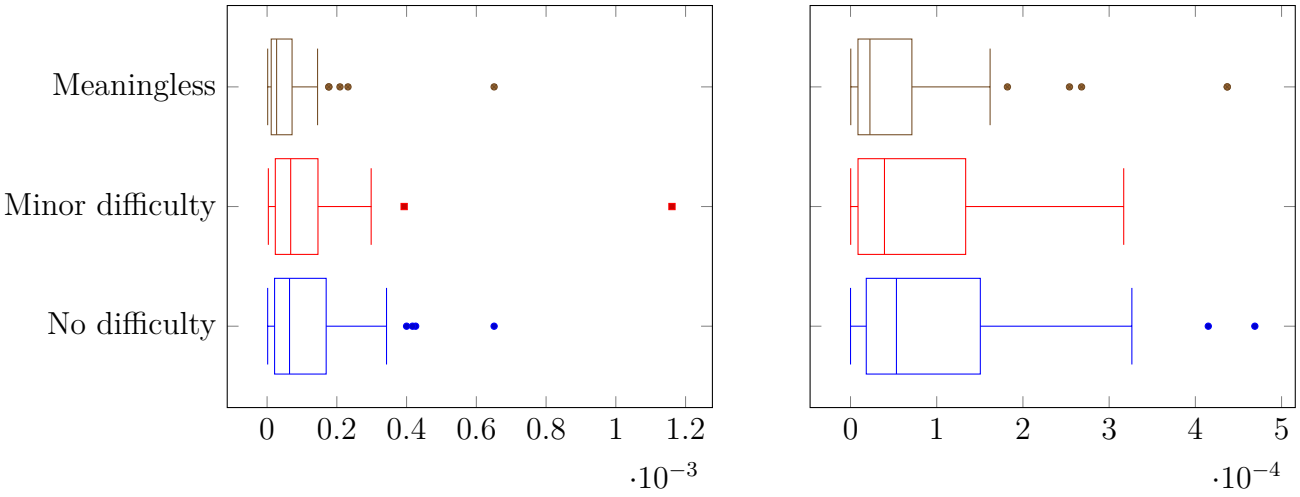


Figure 9: Frequency counts for the heads (left) and modifiers (right) of compounds, where interpretability labels were determined by taking a majority vote over. Whiskers are calculated as the highest or lowest data points within $1.5 * \text{IQR}$ of each quartile, and outliers are depicted with circular or rectangular marks.

for the three box plots, for both heads and modifiers, are clustered together.

While the plots in Figure 9 do suggest that there was a relationship between interpretability label and frequency of occurrence, the two observations above make it clear that there were both: (1) compounds composed of very common words that received *Meaningless* labels, and (2) compounds with very uncommon words that received *No difficulty* labels.

These observations hold true even if we plot the minimum frequency of any word in a compound, rather than treating the frequencies of the heads and modifiers separately. In other words, for each compound, we compute its frequency as by taking the lesser of the frequencies corresponding to its head and modifier. The box plot for this minimum-frequency approach is presented in Figure 10, which again contains a separate box plot for each of the three interpretability labels, as determined by majority vote.

In Figure 10, we again see a large set of *Meaningless* compounds with frequently-occurring heads and modifiers, as well as a clustering near the low end of the frequency scale.

From Figures 9 and 10, we can see a clear correlation between frequency of words and interpretability labels. This correlation is somewhat unfortunate but reflects the reality of an experimental setting in which human judges often treat uncommon words as more 'difficult' to interpret. However, in the analysis above, we see that the frequency of words was not a crucial factor in determining interpretability labels, as many compounds based on common words were labeled as *Meaningless*, and compounds based on rare words were often labeled as interpretable with *No difficulty*.

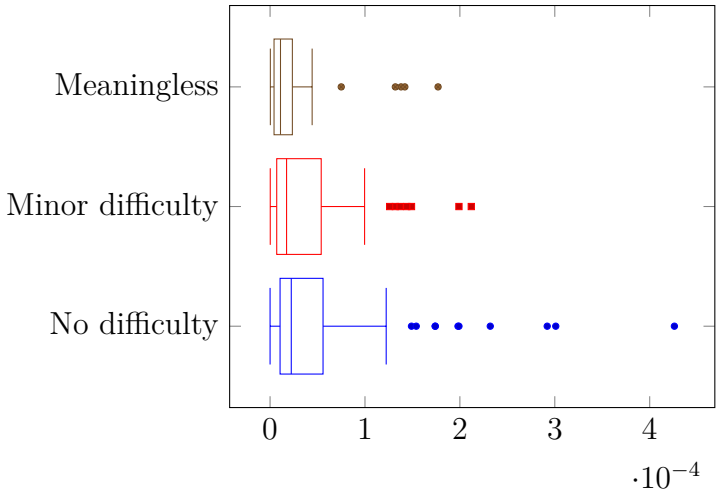


Figure 10: Frequency counts for the minimum frequency of a compound's head and modifiers, where interpretability labels were determined by taking a majority vote. Whiskers are calculated as in Figure 9.

6.5 Clustering Over Paraphrase Dependency Representations

In this section, we explore the impact of noun compound interpretability on paraphrase structure. In particular, we construct vector-space models over dependency representations of the paraphrases collected on the AMT platform, and use these models to establish relationships between paraphrase structure and ease of interpretation.

In linguistic analysis, it is common to convert documents from raw text to vector-space representations. Typically, a document will be conceptually represented using a vector of keywords, with each of those keywords being assigned a weight to represent its importance both to the document and within the collection of documents. These output representations are often easier to work with and allow for computations that would not have been possible had the documents been represented as raw text [22].

As an example, a simplest model could count the number of occurrences of each word in each document. Document similarity could then be computed by measuring the pairwise cosine similarity of each vector. Even with this simplistic model, one can gain deep insights into the structure of documents.

Common tasks made easy through the use of vector-space representations include:

- Determining which documents are 'similar' in a set of documents [12].
- Determining which words are most important in distinguishing the content of a document [36].
- Determining various ways to cluster or categorize sets of documents [12].

In our analysis, we wanted to take a closer look at the paraphrases provided by Turkers and, in particular, answer the above questions by treating our paraphrases as documents. For

example, we may wish to identify the distinguishing factors of paraphrases, especially those for which human judges had difficulty devising a reasonable interpretation. Answering these questions would put us one step closer towards our goal of understanding the interpretability of compounds.

However, it is important to recognize that our interest did not lie in the actual *content* of the paraphrase. Given that we typically only examined the core of the paraphrase (i.e., the blank that Turkers were asked to fill in), most paraphrases did not contain a significant number of tokens; often, the only non-preposition or article would be a lone verb, and deriving insights on which verbs Turkers used would tell us more about the actual interpretations of the compounds, rather than the process of paraphrasing.

Instead of using the paraphrases directly, then, we first passed each paraphrase through a *dependency parser* and used the output representation as its representative document. These representations captured the underlying grammatical and semantic structure of paraphrases, rather than the exact words that composed them, which were relatively insignificant in comparison.

As an example, given the compound *surface colonies*, a human judge could submit the paraphrase, "a surface colony is a colony of people on a planet's surface". Had we used the actual paraphrase as the representative document, our model would focus on keywords, like 'people' and 'plant'. But when comparing this paraphrase to others, what's more important is its underlying structure, including the use of multiple propositions and a possessive term.

After computing the dependency representation of each paraphrase, we ran two models over the set of documents:

- *Term Frequency-Inverse Document Frequency (TF-IDF) Indexing*: The TF-IDF model computes the number of occurrences of each word across a set of documents and separates out commonly-used words (especially stopwords, like 'the') from rarer, more substantive words (like 'healthcare' or 'warfare'). The TF-IDF model assigns high weights to words that appear frequently, but only in a small number of documents within the overall set, as these words are likely indicative of the content of said documents [22].

- *Latent Dirichlet Allocation (LDA)*: LDA centers on a creating a generative probabilistic model of a corpus of documents. It is best known for its use in topic modeling, the task of grouping documents based on common topics, which typically involves identifying which topics influence which documents, and in what proportions [5]. For example, with LDA, one might be able to discern ten different topics governing a set of journal articles, which could range from biology to computer science. An article on computational biology, then, would be influenced by both the biology and computer science topics, in some quantifiable degree.

In the sections that follow, we will present the results produced by these two models. While this analysis does not directly focus on the question of what makes a compound interpretable, developing a more complete understanding of paraphrase construction is key to composing a comprehensive theory of interpretability.

The Stanford Dependency Representation

Dependency parsing is the task of uncovering the grammatical relationships between words in a sentence. For example, given the sentence "The baby is cute", an accurate dependency parser would reveal that the 'baby' is the nominal subject of the clause.

There are a number of possible representations through which to express such dependencies. Our representation of choice was the Stanford Typed Dependencies (SD) representation, a representation "designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used." The SD representation includes approximately[10] 50 different grammatical relations, each of which, significantly, is expressed as a binary relation between two tokens. The *numeric modifier* grammatical relation, for example, is used when a number phrase modifies a noun by associating it with a quantity, like "Charlie spent ten dollars". In this case, the *numeric modifier* relation would be between the words 'ten' and 'dollars'. In this way, each grammatical relation in the output representation captures the connection between two tokens [7].

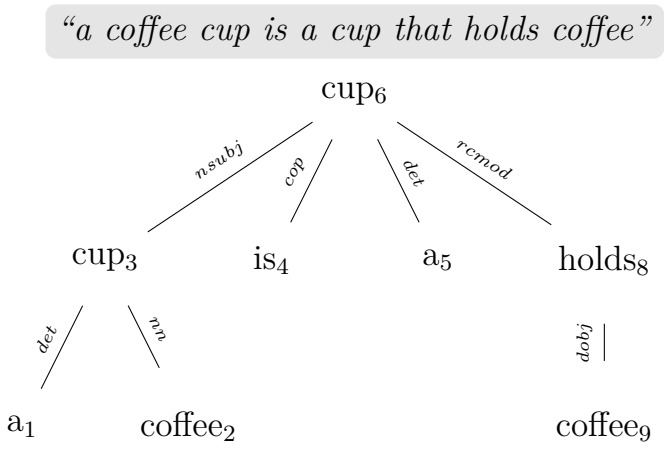


Figure 11: The paraphrase "a coffee cup is a cup that holds coffee", decomposed using the Stanford Typed Dependencies (SD) representation, where subscripts refer to the position at which each word occurs in the paraphrase. Reliance on the relative clause is reflected in the link between the cup_6 and $holds_8$ nodes, which are connected with the relative clause modifier (*rcmod*) binary relation.

The SD representation is useful in that it is simple, expressive, and easy to understand, even to those without a strong background in linguistics. As an added bonus, the Stanford Parser, an open-source tool, is freely available to quickly and accurately compute SD representations of input phrases. As such, we used the Stanford Parser in the analysis that follows. Specifically, we used v3.5.0 of the probabilistic content-free grammar (PCFG) parser implemented by Klein and Manning [18] and trained over the SD representation [8].

[10]The approximation comes from the fact that, based on configuration parameters, the parser can vary the set of possible relations.

A sample compound, along with a paraphrase and its respective SD representation, is presented in Figure 11.

TF-IDF Indexing

Our first approach was to run the TF-IDF algorithm over the SD representations of our paraphrases, which, in the terminology of traditional vector-space model-based techniques, served as our documents.

For each paraphrase, we first cleaned it through the process described in Section 6.2. The cleaned paraphrase was then run through the Stanford Parser [18]. The output representation from the Stanford Parser is a list of binary representations between words. The raw relations were extracted (e.g., `nn(coffee, cup)` was transformed to `nn`), and these relations served as the tokens for our TF-IDF model. Note that the TF-IDF model, as with most vector-space models, uses a bag-of-words representation in which the ordering of tokens in a document is considered irrelevant [22]. As such, the fact that these relations were between tokens in the sentence that might not be adjacent, or that the relations may be out of order, is irrelevant, given that these models only rely on frequency counts.

As an example, consider the compound and paraphrase from Figure 11. The SD representation of this paraphrase was as follows:

```
det(cup-3, a-1)
nn(cup-3, coffee-2)
nsubj(cup-6, cup-3)
cop(cup-6, is-4)
det(cup-6, a-5)
root(ROOT-0, cup-6)
nsubj(holds-8, cup-6)
rcmod(cup-6, holds-8)
dobj(holds-8, coffee-9)
```

By stripping away the words in the relation and considering this to be an unordered list of tokens, this paraphrase was represented with the vector: `[det, nn, nsubj, cop, det, root, nsubj, rcmod, dobj]`.

After computing these token vectors, we ran them through the TF-IDF Vectorizer implementation provided by the `scikit-learn` Python library [35]. This produced an output matrix M in which each row represented a document, each column represented a token, and entry $M(i, j)$ represented the importance of token j to document i, computed as the product of the frequency of the token in the document and the number of documents containing that token.

Once this matrix had been computed, there were several interesting operations that we could perform, such as:

- *Computing the paraphrases that were most similar (at a structural level) to a given paraphrase.* This was calculated by taking the cosine similarity of the TF-IDF vectors

for each document, a useful metric for determining document similarity given vector-space models [14], and reporting the n most similar documents (i.e., the n documents with the highest cosine similarity score).

- *Computing clusters of similar paraphrases.* This was done by running the k-Means algorithm over the TF-IDF vectors. k-Means, an iterative vector clustering technique, aims to find the k 'center' points that minimize the cumulative squared distance from any vector to its nearest center. The algorithm thus provides a method for grouping documents by their respective cluster assignments. In particular, a document's cluster is determined by the center point to which it is closest [16]. In terms of paraphrases, each cluster could be seen as representing a unique paraphrase structure.

Some sample results for the first task (computing similar paraphrases) are presented in Figure 12, which displays the three most similar paraphrases for a number of different paraphrases. The key observation is that the structure captured by the TF-IDF vectors extends beyond the mere identification of shared prepositions. Instead, we see how those prepositions are used (e.g., in front of a verb vs. standing alone).

an enterprise product is a product that is maintained by an enterprise
a cathedral performance is a performance that was performed by the cathedral
a government statue is a statue that was erected by the government
a cathedral performance is a performance that is given by a cathedral

a sugar measure is a measure of sugar
an oil ring is a ring of oil
a cotton order is an order of cotton
an accident dispute is a dispute of an accident

a hotel model is a model that is a spokesmodel for a hotel
a business party is a party that is for the business
a city engineer is an engineer that works for a city
a city engineer is an engineer that works for the city

a pressure dispute is a dispute that is under a lot of pressure
a career practice is a practice that is part of a career
a player industry is an industry that comprise of player
a machine core is a core that is a part of the machine

Figure 12: The three most similar paraphrases for several different target paraphrases. The similarities reveal key structural elements of the paraphrases beyond their preposition of choice. For example, the top-left paraphrases share the use of a verb-'by' pattern. The top-right paraphrases are concise, using only the 'of' preposition; however, this is in stark contrast to the bottom-right paraphrases, which also use the 'of' preposition, but with a leading verb.

For the second task (clustering paraphrases), results are contingent on the number of clusters that we choose to compute, i.e., the choice of k in the k-Means algorithm. For the sake of demonstration, we choose $k = 8$. The results of the k-Means algorithm also vary based on a chosen random seed. We set the global random seed to 0 using the `np.random.seed(seed=0)` function provided by the Numpy scientific computing library [42]. The output clusters are displayed in Figure 13 below.

From Figure 13, the common structure of paraphrases, as grouped by cluster, is evident immediately. For example, some clusters, like `Cluster 0` and `Cluster 7`, contain very simple paraphrases based on prepositions. `Cluster 1` again contains simple paraphrases, but with a reliance on verbs. Meanwhile, `Cluster 4` and `Cluster 5` contain more complex

paraphrases involving a range of grammatical structures, which are themselves indicative of more elaborate explanations.

`Cluster 0`
drug orders are orders for drug
a party soup is a soup for a party
a peer version is a version that is for peers

`Cluster 1`
a life zone is a zone that protects life
a city dispute is a dispute that concerns a city
a nut engineer is an engineer that makes nuts

`Cluster 2`
government bars are bars that are run by the government
a student paper is a paper that was written by students
a chocolate burn is a burn that is caused by hot chocolate

`Cluster 3`
city members are members that live in the city
a neighborhood lake is a lake in a neighborhood
a future actor is an actor that will act in future

`Cluster 4`
blanket months are months of the year when you need a blanket to stay warm
an automobile dune is a dune that is to be driven on by automobiles
a part decision is a decision on who will be playing which part

`Cluster 5`
a pressure dispute is a dispute that is under a lot of pressure
a sea machine is a machine for converting the motion of the sea to energy
citizen teams are teams that are made up of citizens

`Cluster 6`
a top group is a group that is on top
surface colonies are colonies on contaminated surface
an acting fair is a fair where people showcase their acting

`Cluster 7`
a bacon sauce is a sauce from bacon
enemy signals are signals from the enemy
a sports price is a price that you place on sports

Figure 13: Sample paraphrases from each of the eight clusters generated by the TF-IDF method. The distinctive characteristics of each cluster are evident, even with the small sample size. For example, `Cluster 0` contains very simple paraphrases that make use of prepositions, especially "for"; `Cluster 1` too contains simple paraphrases, but in this case, these paraphrases rely on verbs; on the other hand, `Cluster 4` contains complex, diverse paraphrases involving a wide range of grammatical structures.

We can use these clusters to identify structural differences between the paraphrases associated with *Minor difficulty* and *No difficulty* judgments. In Section 6.2, we saw that compounds that are more difficult to interpret were given longer, more complex paraphrases by human judges. Now that we have developed a framework for clustering paraphrases based on grammatical and structural properties, we can try to extend this observation into a more nuanced model. In particular, for each of the k clusters output by the k-Means algorithm, we can check what percentage of the paraphrases in that cluster belonged to *No difficulty*, *Minor difficulty*, and *Eccentric* judgments.

Note that, for this analysis, we're no longer determining the appropriate interpretability label for a given paraphrase by taking a majority vote over the judgments leveled on its matching compound. Since this analysis is per-paraphrase, we can instead use the interpretability label associated with that paraphrase directly.[11] For that reason, each *Eccentric* judgments would also qualify as a *No difficulty* or *Minor difficulty* judgment, based on the definition of *Eccentric*. To avoid over-counting, we thus treat *Eccentric* paraphrases as solely *Eccentric*, and ignore the fact that the user labeled it as interpretable with *No difficulty* or *Minor difficulty*.

[11]For example, under a majority vote scheme, if a compound received two *Minor difficulty* judgments and one *No difficulty* judgment, we would still label each paraphrase as paired with a *Minor difficulty* compound. However, in this section, we would pair each paraphrase with the interpretability label of its judgment directly.

The results of this analysis are presented in Figure 14. To highlight the difference in breakdowns across clusters, we normalize the counts by dividing the number of paraphrases of a certain type in each cluster by the total number of paraphrases in that cluster.

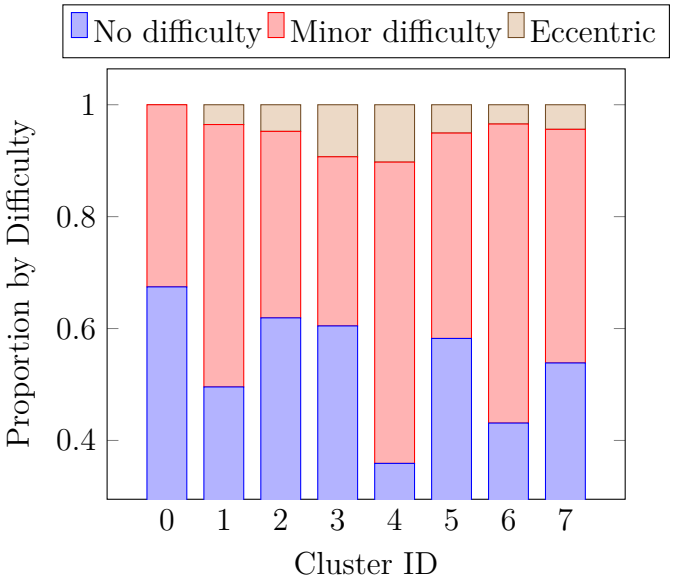


Figure 14: A breakdown of the paraphrases in each cluster by difficulty of interpretation. `Cluster 0`, with no *Eccentric* paraphrases and a larger proportion of *No difficulty* paraphrases than any other cluster, seems to capture some grammatical or semantic structure indicative of ease of interpretation.

Certain clusters, like `Cluster 4`, seem to identify paraphrases linked to difficult interpretations, as the normalized counts indicate that this cluster contained a larger proportion of *Minor difficulty* and *Eccentric* paraphrases than any other cluster. On the other end of the spectrum is `Cluster 0`, which contains more than twice as many *No difficulty* as *Minor difficulty* paraphrases–and no *Eccentric* paraphrases. If we look at the contents of these two clusters, `Cluster 0` contains paraphrases that are short and preposition-based, like *"drug orders are orders for drug"*, while `Cluster 4` contains paraphrases that are longer and, for lack of a better word, more eccentric, utilizing grammatical devices that are relatively uncommon. A useful example comes from the compound *wastebasket inventions*, for which one judge submitted the following paraphrase: "wastebasket inventions are inventions that no [sic] practical and balled up and thrown away in wastebasket".

While these observations are somewhat anecdotal and based on just one run of the k-Means algorithm, the results are promising. In particular, they demonstrate that vector-space models can be useful in developing a better understanding of how paraphrases are constructed and the extent to which a paraphrase's structure reveals information about the difficulty of its composition. Importantly, these results also validate hypothesis **H5** from Section 3, which claimed that more complicated paraphrases would correlate with difficulty of interpretation.

Latent Dirichlet Allocation

Next, we ran a similar experiment to that described in the previous section, but with a very different model anchoring our exploration: the Latent Dirichlet Allocation (LDA) model of Blei et al. [5]. As discussed above, LDA takes a Bayesian approach by constructing a generative probabilistic model that views documents as mixtures over a set of underlying topics. Topics, in turn, are modeled as distributions over tokens.

In general, LDA attempts to improve over the TF-IDF model by providing a better measure for inter- and intra-document statistical structure, per Blei et al. [5]. As with the TF-IDF model, topics in the LDA model are typically computed over a document's raw text tokens, rather than other lexical features, like part-of-speech tags or, in our case, binary dependency relations between tokens in the document. However, such an approach is not unprecedented. For example, Wong et al. [44] showed that topic modeling with LDA over part-of-speech tags is useful in solving the problem of native language identification. Informally, Wong et al. found that certain bigrams of part-of-speech tags were common in certain languages, which rendered them useful in determining a speaker's native language. In fact, according to Wong et al. [44] their topics over part-of-speech tags provided more useful clustering of terms than their topics of functional keywords.

It is under a similar motivation that we bring topic modeling, and the LDA model in particular, to the analysis of paraphrases. Specifically, and as in the TF-IDF experiment above, we ran our paraphrases through the Stanford Parser and subsequently constructed a topic model over the paraphrases' SD representations. In this topic model, topics composed mixtures over SD binary relations; as such, each topic could be viewed as representing a grammatical style.

Our experiment relied on the `lda` package hosted on the Python Package Index. The `lda` implementation is based on collapsed Gibbs sampling. Additionally, as LDA is a technique that requires the number of underlying topics to be hard-coded and decided upon beforehand, we present in this section results from just one run of the algorithm, with the goal of presenting a glimpse into the potential usefulness of topic modeling and, as with the previous section, vector-space models for understanding paraphrases.

As with the TF-IDF-based analysis, we were required to determine the number of topics in advance. In this case, we chose to construct five distinct topics. Again, randomization was standardized by calling the `np.random.seed(seed=0)` function provided by Numpy [42]. A portion of our results are displayed in Figure 15 which contains, for each of the five topics:

- The three most frequent SD relations for that topic, ignoring `det`, `nsubj`, `nn`, `root`, `cop`, and `rcmod`, as these relations were present in nearly every paraphrase due to the structure enforced by our HIT template, as described in Section 4.2.

- Two paraphrases for which that topic contributes more than any other topic, based on the underlying mixture. These paraphrases are thus considered to be representative of the style captured by that topic.

The results produced by the LDA model differ from those of the TF-IDF model described in the previous section. While the divergence in output speaks, at some level, to the differ-

ences between the two models (which are not entirely relevant to this discussion), in terms of the usefulness of analyzing paraphrases, we can continue to learn about the distinctive elements of paraphrase structure. In general, it was more difficult to assign clear interpretations to the LDA topic clusters, perhaps due to the lack of discrimination between frequently-used relations and those that capture important characteristics of a paraphrase.

The most interesting topic is `Topic 0`, which appears to capture the passive voice, making strong use of the `nsubjpass`, or *passive nominal subject*, and `auxpass`, or *passive auxiliary* relations. By segmenting the paraphrases based on their most influential topics, we also find that `Topic 0` is indicative of low difficulty paraphrases. In particular, of the paraphrases for which `Topic 0` was most influential, 62.5% of them were labeled *No difficulty*, 37.5% *Minor difficulty*, and 5.5% *Eccentric*.[12] This was the highest percentage of *No difficulty* and lowest percentage of *Minor difficulty* paraphrases for any topic, by margins of over 4% in both cases. In a sense, then, we can see how topics are a useful mechanism for unveiling hidden grammatical structure, and that this structure can be indicative of the difficulty of interpretation and paraphrasing.

`Topic 0` | `nsubjpass`, `auxpass`, `agent`
trust money is money that has been put in a trust
a jungle paper is a paper that is printed with a jungle scene

`Topic 1` | `dobj`, `prep_for`, `prep_in`
a summer dispute is a dispute that takes place in summer
an automobile dune is a dune that is suitable for automobiles

`Topic 2` | `dobj`, `prep_of`, `amod`
a life zone is a zone that encompasses all aspects of daily life
a daisy baby is a baby that has blonde hair like a daisy

`Topic 3` | `prep_of`, `amod`, `conj_or`
a top group is a group that or the best group or the elite group
an air zone is a zone pertaining to air combat or atmospheric condition

`Topic 4` | `amod`, `dobj`, `aux`
a product step is a step that has to do with product
a jungle range is a range that is with no clear view jungle

Figure 15: Topics constructed using the Latent Dirichlet Allocation (LDA) model. For each topic, we list: the three Stanford Typed Dependency (SD) relations, ignoring relations that were common to every paraphrase; and two representative paraphrases, to which the listed topic was the most influential.

While some topics exhibit distinctive characteristics (e.g., `Topic 3` captures the use of the `conj_or`, or *conjugate 'or'* relation), others overlap or fail to capture an immediately obvious grammatical structure (e.g., `Topic 4` and `Topic 3` both make heavy use of the `amod`, or *adjectival modifier* relation; it is thus difficult to identify either of them as the '`amod` topic').

While this section gave just a glimpse into the use of vector-space clustering for analyzing paraphrase structure, it is clear from the analysis above that the structure of paraphrases can be useful in identifying aspects of interpretative difficulty, and, further, that dependency relations can appropriately capture grammatical structure. However, given the relatively small size of our dataset and the limited nature of this analysis, many of the topics output by our model appear to be lacking in distinctive characteristics. While this could be a function of the number of topics chosen, in the future, it would be interesting to extend this

[12]Recall that an *Eccentric* paraphrase will also be a *No difficulty* or *Minor difficulty* paraphrase, so the three categories do not add up to 100%.

analysis over a larger dataset and dive more deeply into the characteristics that make topics identifiable.

6.6 Training a Classifier

In this section, we explore the task of training machine learning classifiers to identify the interpretability label of a given compound. In particular, we explore two possible approaches and data sources, with the goal of demonstrating that both are useful features in the context of training a classifier and, as such, capturing the notion of interpretability:

1. Comparisons to attested compounds with a shared head (*modifier variants*) or modifier (*head variants*).

2. Paraphrases, as submitted by human judges.

In both cases, we had to develop a formulation in the classical machine learning setting (e.g., determining feature vectors, labels, training and testing data, and so forth). When evaluating the performance, the goal was not to prove that the construction of such a model is *possible*, but rather, to further demonstrate that these sources of data are helpful in assessing the noun compound interpretability.

We start by discussing a formulation of the machine learning problem in which we train our model using semantic similarity metrics and, in particular, comparisons to attested compounds. Afterwards, we present a different formulation in which we rely on user-submitted paraphrases, specifically the TF-IDF vectors based on Stanford Typed Dependency (SD) representations of the paraphrases, as discussed in Section 6.5.

Training Against Attested Compounds

When training a model using comparisons to attested compounds, our basic unit of data is a pair of generated and attested compounds. For example, the pair *cotton cup*, *coffee cup* might produce one such unit of data. However, for each sense (based on WordNet synsets) in which head and modifier of the generated compound was used, we add a separate training example, such that if one judge interpreted 'cup' as a container for holding liquid and another as a trophy, these two senses would merit inclusion as separate examples.

In simplest terms, for each compound, for each judgment,[13], for each attested variant, we add a training example. The training example has a corresponding feature vector that is computed using the synset annotations described in Section 6.1.

To be as explicit as possible, let's walk through a more detailed example, again based on the compound *cotton cup*. Assume this compound has two modifier variants, *coffee cup* and *plastic cup*. When gathering our three judgments on the AMT platform, two judges interpret the modifier *cotton* in a sense best matched to the WordNet synset *cotton.n.1*, while the third

[13]As our feature vectors are computed using WordNet synsets, we have to treat each judgment for a given compound as a separate training example, as different judgments might make use of different WordNet synsets.

judge interprets it in a sense best matched to *cotton.n.2*. Recall that for attested variants, WordNet synsets were determined with the first-sense heuristic, such that the modifier of *coffee cup* would always be represented with the synset *coffee.n.1*. In augmenting our training data set, we would produce a training example based on the following pairings of synsets: *cotton.n.1* and *coffee.n.1*; *cotton.n.1* and *coffee.n.1* again, as *cotton.n.1* was used twice in judgments submitted by Turkers; and, finally, *cotton.n.2* and *coffee.n.1*. A similar set of judgments would be added for the remaining modifier variant, *plastic cup*, and the process would be repeated for head variants.

However, this explanation leaves several questions unanswered, including:

1. How should we decide on a training label for each unit of data? For example, if one compound receives two *No difficulty* judgments and one *Minor difficulty* judgment, do we label every example involving this compound as *No difficulty* (i.e., take a majority vote), or should the data corresponding to the *Minor difficulty* judgment have a *Minor difficulty* label?

2. Which features should we include in the feature vector? Should we stick to WordNet-based semantic similarity measures? Or should we include additional features, like the Latent Semantic Analysis (LSA) similarity score computed between the generated and attested compounds?

3. Should we include an example for *every* pair of generated compounds and attested variants? Or can we come up with a more nuanced approach?

4. Should we use head or modifier variants (as defined in Section 6.3)?

5. If we're relying on WordNet synsets to compute feature vectors, how can we train on *Major difficulty* compounds, for which there is no synset?

In the analysis that follows, we consider each of these questions to be a different configuration parameter and evaluate the results across every combination of parameters. Specifically:

1. When parameter `M` is enabled, we use a majority vote to determine interpretability labels; otherwise, we use the label of the submitted judgment. For example, if a compound received two votes for *No difficulty* and one vote for *Minor difficulty*, we would label every example based on that compound as *No difficulty*, even those examples produced by the *Minor difficulty* judgment.

2. When parameter `F` is enabled, we include two semantic similarity measures that come from outside of WordNet: *Latent Dirichlet Allocation (LDA)* similarity [5], and *Latent Semantic Analysis (LSA)* similarity [19]. Both of these measures aim to capture the similarity of two words based on contextual usage. As such, the output scores are relative to a corpus of text on which a model must be trained. To compute the scores, we used the open-source `SEMILAR` software package from Rus et al. [39], which is trained over the SIMILAR corpus, a combination of the TASA corpus and Wikipedia

[38]. When the parameter `F` is disabled, feature vectors are restricted to the following WordNet semantic similarity measures, computed over the relevant synsets for the given judgment and attested compound: Shortest-Path Distance, Path Similarity, Leacock-Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity, JCN Similarity, and Lin Similarity.[14]

3. When parameter `C` is enabled, for each compound, we take the set of synsets that map to all of its attested variants, add in the synset for the current judgment, and run a graph clustering algorithm over the synsets (specifically, the Affinity Propagation algorithm from Frey and Dueck [11], which allows for unsupervised clustering and automatically discerns an appropriate number of clusters); we then limit the included examples to those for which the attested synset ended up in the same cluster as the generate compound's synset. The intuition here is that a set of attested compounds may represent a variety of interpretations of the shared head or modifier, and by clustering, we can instead compare the generated compound to the set of attested compounds to which it is semantically closest, thus reducing noise.[15] If `C` is disabled, all pairs of generated compounds and attested variants pairs are included, indiscriminately.

4. When `variant=head`, experiments used head variants; when `variant=modifier`, modifier variants were used instead. As such, experiments were run separately on head and modifier variants, with the aim of determining which component (i.e., the head or the modifier) is most useful in assessing interpretability.

5. *Meaningless* compounds were included in the dataset, and the first-sense heuristic was used to assign synsets to their heads and modifiers. Since many of the questions around interpretability involve determining whether a compound is interpretable *at all*, it was considered important to include these *Meaningless* compounds.

For each permutation of parameters, we trained an AdaBoost classifier with decision trees. To be precise, ensembling was limited to 100 or 1,000 weak learners (the exact choice is made clear when reporting results), which were themselves limited to a depth of two. Performance was assessed using k-fold cross validation with $k = 8$, with accuracy rates averaged across folds.

In Table 9 and Table 10, we present the results for a classifier trained on comparisons to head and modifier variants, respectively. Both tables include separate accuracy scores for each combination of configuration parameters.[16] In addition, we note the baseline accuracy

[14]For those WordNet-based similarity metrics that relied on a measure of Information Content, the Brown Corpus was used due to ease of implementation. However, it is noted that these measures could be improved by using a larger and more substantive corpus.

[15]As an example, assume the modifier of *cotton cup* is interpreted by one judge using the synset *cotton.n.1*. The attested modifier variants include *coffee cup* and *plastic cup*, which are then represented with the synsets *coffee.n.1* and *plastic.n.01*. When clustering over the WordNet graph, *plastic.n.1* and *cotton.n.1* are placed in the same cluster as they are both materials, but *coffee.n.1* is placed in a separate cluster. If `C` were enabled, only the variant *plastic cup* would be considered when adding training examples.

[16]Certain superfluous combinations of parameters are excluded.

(i.e., the frequency of the most commonly occurring label) for every such combination, which varied as certain parameters impacted the manner in which examples were generated and the criteria by which they were included.

`Head Variant Comparisons`

Num. Learners	M + C + E	M + C	M	C	(None)
Baseline	46.18	46.18	65.28	41.99	56.71
100	74.66	71.88	66.47	54.87	57.28
1,000	84.08	80.89	70.57	58.53	57.45

Table 9: The accuracy of an AdaBoost classifier with two-level decision trees, trained on comparisons between generated compounds and attested *head variants*, over a variety of configuration parameters. The values in the table represent percentages, computed as an average over eight folds in a k-fold cross validation setup.

`Modifier Variant Comparisons`

Num. Learners	M + C + E	M + C	M	C	(None)
Baseline	46.30	46.30	49.14	45.51	47.52
100	74.99	69.34	57.81	53.19	49.89
1,000	86.28	81.70	66.91	56.18	53.16

Table 10: The accuracy of an AdaBoost classifier with two-level decision trees, trained on comparisons between generated compounds and attested *modifier variants*, over a variety of configuration parameters. The values in the table represent percentages, computed as an average over eight folds in a k-fold cross validation setup.

There's much to be learned from the results in these two tables. When comparing against both head and modifier variants, *learning against majority-voted labels was far easier and more successful* than learning against per-judgment labels. In the latter scenario, it was possible to have multiple identical training examples with different interpretability labels,[17] which led to contradictory data and an inherently impossible learning task; thus, it is not surprising that the majority formulation was more successful.

Additionally, *we found that clustering attested variants was a highly effective technique* for improving performance. For head variants, the use of clustering led to accuracy increases of over 5% and 10% for 100 and 1,000 learners (against a baseline that was nearly 20% more difficult), while for modifier variants, accuracy increased by nearly 12% and 15% for 100 and

[17]For example, if the compound *cotton cup* was interpreted by one judge as interpretable with *No difficulty* and, by another, as interpretable with *Minor difficulty*, and assuming both judges used the sense *cotton.n.1* (as identified by the content of the paraphrases they submitted), and assuming that the modifier variant *coffee cup* is present in our attested dataset, we would include the examples (f(*cotton.n.1*, *coffee.n.1*), *No difficulty*) and (f(*cotton.n.1*, *coffee.n.1*), *Minor difficulty*). Thus, two examples would be produced with an identical feature vector and non-identical labels.

1,000 learners (against a baseline that was around 2% more difficult). This speaks to the suggestion, proposed earlier, that attested variants might represent a few different, broad senses, and that honing in on the closest cluster of senses would be a useful technique. In other word, the results above would suggest that the *closest* cluster of semantically similar variants is far more useful in judging interpretability than, say, the average variant.

Further, *the use of extra, non-WordNet-based features (in this case, LDA and LSA similarities) led to improved performance*, typically increasing accuracy by around 5%, a number that was consistent between head and modifier variants, and across the number of weak learners. The usefulness of these features is further evidence that these types of similarity metrics (in this case, semantic measures based on contextual similarity) can help capture the notion of interpretability for new compounds, especially insofar as they relate to attested compounds.

It should be noted, however, that this formulation of the task is not completely faithful to reality. Specifically, for each judgment of a generated compound, for each attested variant, we're adding a unit of training data to the dataset. As such, our results could be biased towards compounds with many attested variants or be otherwise skewed in a way that is not immediately obvious.

In a more realistic formulation, one might represent each generated compound (or each judgment) as a single unit of data, computing a single feature vector to capture the similarity of that judgment to every attested variant. For completeness, we implemented a scheme that followed this logic. In particular, we left the data generation step as above, but when computing output labels over the test set, took a majority vote over every unit of data linked to a given generated compound.

However, in this formulation, classifiers performed poorly. The most significant issue was a lack of data: between folds in the k-fold cross validation, accuracy fluctuated from 20% to 60%, as the size of our dataset was reduced to fewer than 250 data points (i.e., the number of compounds for which we have a clear majority label), whereas in the previous setup, our dataset contained over 14,000 (generated, attested) pairs.

As a whole, the complicated and nuanced formulation of this experiment left us skeptical of its significance–or, at the very least, the applicability of its results. In the end, this analysis was most useful for identifying factors that are important or unimportant in determining the interpretability of a compound, e.g., the use of LSA and LDA, and the clustering of attested variants. Further, the machine learning techniques explored in this section represent just a small sampling of those used across the field. Future work could focus on testing a wider range of classifiers trained on a larger dataset, which would allow for a formulation of the problem in a manner more suitable to real-world applicability (i.e., in which each compound is represented by a single training example).

Training Against Paraphrases

In this section, we explore the idea of training a machine learning classifier using the TF-IDF vectors from Section 6.5. The setup here is far simpler than that of the attested variant comparison-based classifier described above. Specifically, as each judgment includes both

an interpretability label and a paraphrase, we map judgments to training examples in a one-to-one manner by treating the interpretability label as a training label and the TF-IDF vector (computed over the SD representation of the paraphrase) as its feature vector.

As such, our features are TF-IDF frequencies of the various binary relations used in the SD dependency representation system. In this formulation, then, producing a classifier that outperformed the baseline would imply that the grammatical structure of a paraphrase helps capture the interpretability of its corresponding compound. A similar claim was made in Section 6.5; we attempt to bolster it further with evidence drawn from the machine learning setting.

Before presenting the results produced by classifier, we first discuss the details of our training and testing environments. These are introduced via comparisons to the setup of the previous section:

- In the previous section, we took the majority vote over a compound's labels as the ground-truth label for our machine learning classifier. Part of the reason that this was necessary was that we were adding a different training example for each judgment and for each attested variant. Thus, if two judgments (for the same compound) differed in their interpretability labels and we did *not* use the majority vote as its label, we could end up with contradictory training data in the form of identical feature vectors with non-identical labels. In our new setting, this concern is no longer relevant, as no two paraphrases for the same compound are (presumably) similar enough to lead to the same feature vector. *As such, for each judgment submitted by Turkers, we use the interpretability label of that judgment (rather than a majority vote) when training our model.*

- In the previous section, recall that we were able to include *Meaningless* compounds by using the first-sense heuristic to guess the most relevant WordNet synset for that compound. As a reminder, this measure was required due to the lack of a paraphrase for *Meaningless* judgments; compounds that are deemed *Meaningless* by Turkers, by definition, should not be paraphrasable. As the setup in this section relies so strongly on the presence of paraphrases, we cannot gloss over their absences in the same way. *Thus,* Meaningless *compounds had to be excluded from our analysis, making this a two-label problem in which the classifier was trying to discern whether a paraphrase corresponded to a* No difficulty *or* Minor difficulty *judgment.*

Given these decisions, we ended up with 545 training examples that followed the split presented in Table 11.

These 545 examples were fed to a Support Vector Machines (SVM) classifier using the radial basis function (RBF) kernel, a regularization parameter of $C = 1.0$, a gamma parameter of $\gamma = 0.001$, and auto class weighting. These parameters were selected through a grid search over three different kernels (RBF, as well as the sigmoid and linear kernels) and an exponential range of regularization and gamma values. The accuracy of a classifier trained on each combination of parameters was assessed using k-fold cross validation for $k = 8$.

Category	Size
No difficulty	302
Minor difficulty	243
Total	545

Table 11: Breakdown of the training dataset for the paraphrase-based machine learning classifier, where the value in the right column indicates the number of examples falling under the designation listed on the left.

After tuning these parameters, we evaluated our model using k-fold cross validation for $k = 10$, averaging the accuracies reported over 10 random restarts. In effect, we ran the 10-fold cross validation 10 separate times, which allowed us to evaluate performance over 100 distinct folds. The results are presented in Table 12.

Round	Folds > Baseline	Avg. Accuracy (%)
1	**5**	**55.63**
2	**6**	**56.68**
3	**6**	**56.51**
4	4	**55.98**
5	**6**	**56.39**
6	**7**	**56.49**
7	**8**	**57.96**
8	**7**	**56.89**
9	**7**	**56.86**
10	**5**	55.24

Table 12: Performance of an SVM classifier trained to predict a judgment's interpretability label (either *No difficulty* or *Minor difficulty*) using a vector-space representation of its corresponding paraphrase. Each round consisted of a 10-fold cross validation assessment. In the middle column, we report the number of folds in that round for which accuracy was strictly greater than the baseline (55.3%) and bold the result if at least half of folds represented an improvement. On the right, we display the average accuracy, which is similarly bolded if it represents an improvement over the baseline.

As is evident from Table 12, average accuracy was strictly greater than the baseline of 55.3% for all but one random restart. Similarly, for all but one random restart, at least half of folds performed above the baseline. Taking the average accuracy across all folds and all random restarts, the classifier correctly predicted the label 56.46% of the time, an increase over the baseline of more than 1%. If we restrict our view to those 61 folds for which we improved over the baseline, the average accuracy was 60.92%, an increase of over 4%.

While these increases may seem minor, it is important to recall that this experiment consisted of just over 500 labeled examples–a relatively small dataset for a machine learning

problem. Additionally, the consistency of these increases across random restarts and random folds is evidence of their significance. In other words, the differences represent more than just random noise.

The feature vectors used in this experiment could be augmented in multiple ways, some of which would likely lead to the training of a more accurate model. However, the goal of section is not to optimize for model accuracy; instead, we aim to validate the usefulness of paraphrases in informing compound interpretability.

Thus, we focus on the existence of this performance increase, rather than its magnitude. And with that in mind, the increase is rather remarkable. Recall that the feature vectors used in this setting did not contain *any* of the original content from the user-submitted paraphrases. Instead, they contained the TF-IDF scores as computed over the dependency representations of those paraphrases. *Using these dependency relationships and nothing else, we were able to train a classifier to predict a judgment's corresponding interpretability label and consistently improve over the baseline.* The conclusion: paraphrase structure–and structure alone–is a *true* and useful indicator of compound interpretability.

7 Extending to Peer Compounds

In our first round of experiments on the AMT platform, we asked human judges to label unfamiliar, machine-generated noun compounds as interpretable with *No difficulty* or *Minor difficulty*, or indicate that they were *Meaningless*, as well as provide a paraphrase for the compound, if possible. The goal of these experiments was to enable us (the experimenters) to develop a theory of noun compound interpretability.

As seen in previous sections (most notably, in Sections 6.3 and 6.6), a driving force behind our analysis was the use of semantic similarity measures, especially those based on WordNet synsets, to evaluate how closely a new, unfamiliar compound reflected those with which human judges were already familiar. In particular, we found that the interpretability of generated compounds correlated with the distances between those compounds and their attested variants, where 'distance' was evaluated on the basis of semantic similarly.

These findings inspired us to explore the idea of generating new compounds based on predictable deviations from the compounds we generated to create our initial dataset. This new batch of generated compounds could then be used to evaluate the effect of these predictable deviations on compound interpretability.

To be more precise: after running our initial batch of experiments and annotating the user-submitted paraphrases with WordNet synsets as described in Section 6.1, this left us with a specific sense in which a modifier and head was used, for each noun compound. By mutating those senses through specific WordNet relations, like those defined in Section 2.4, we were able to create new noun compounds, referred to as *peer compounds*, that differed predictably from the original compounds, known as *root compounds*, from which they were generated. In particular, the peer compounds could be more abstract (if they were generated by moving upwards in the WordNet graph), more specific (if they were generated by moving downwards), or differ in some other way from the root compounds. (The exact methodology

behind the dataset generation is detailed in Section 7.3 below.) Importantly, the root compounds from which these peer compounds were created had already been assigned 'ground truth' interpretability labels based on the data collected in the first round of experiments. This allowed us to make precise observations as to how deviations in the WordNet graph were reflected by changes in interpretability labels.

Once we had generated our batch of peer compounds, we ran the same experiment on the AMT platform as in the previous section (see Section 4.3 and Section 4.2 for details), but over this new dataset. The same rating system for interpretability (based on the scale of *No difficulty*, *Minor difficulty*, and *Meaningless*) was used, and the same format for paraphrases (based on the relative clause or use of a preposition) was requested and required. Upon completion, the same steps were taken to annotate WordNet synsets, clean paraphrases, etc.

This portion of the report is structured as follows: First, in Section 7.1, we discuss some of the motivating factors and ideas behind this experiment; Section 7.2 then translates these motivations into key hypotheses; next, in Section 7.3, we provide details as to how our new dataset was constructed; in Section 7.4, we discuss some of the high-level statistics from the experiment before analyzing the results in more detail in Section 7.5.

7.1 Motivation

We begin with a discussion of the intuition behind this round of experiments.

Recall that, in WordNet, synsets are structured such that more abstract concepts are positioned closer to the WordNet root. By replacing either the head or modifier of a noun compound with a word referring to a synset that is either higher, lower, or at an equivalent level (i.e., one that shares a superordinate or parent node) in the WordNet tree, we can thus produce a new compound with constituent components that are more, less, or equally specific.

For example, given the compound *orange juice*, we could replace *orange* with a more abstract word (like *fruit*) to create a more abstract compound (like *fruit juice*). Alternatively, we could replace *orange* with a more specific word (like *clementine*) to create a more specific compound (like *clementine juice*). Finally, by finding a synset with a common superordinate (like *apple*, which might share the parent *fruit*), we could create a compound that is equally specific but distinct (like *apple juice*).

The Basic Level

In discussing the role that hierarchy-based changes could play in compound interpretability, it is helpful to briefly mention the idea of a *basic level*, as introduced by Rosch et al. [37]. In that seminal work, Rosch argued that, in hierarchical taxonomies of basic objects, there exists one level in the taxonomy, known as the *basic level*, for which "categories carry the most information, possess the highest cue validity, and are, thus, the most differentiated from one another".

In simpler terms, the basic level includes those objects that are considered prototypical of a certain class or category. They are the objects often listed by human judges when asked

to name a "typical" version of that class of objects and, as such, usually represent the terms that occur most often and with which judges are most familiar.

For example, given the category *fruit*, members of the basic level might include *apple*, *banana*, and *grape*, but not *pomegranate* or *guava*. Note that these first three fruits are commonly occurring and highly differentiated; if we move down any of their respective subtrees, the objects became much less so.

Traveling down the *apple* subtree, for example, the next level might contain *Gala apple* and *Mackintosh apple*. These two objects are much less commonly occurring and much less differentiated than, say, *apple* and *banana*, or *apple* and *grape*. Similar observations can be made for other superordinates, like *tool*, where the basic level might include *hammer*, *saw*, and *screwdriver*; and *furniture*, where the basic level might include *table*, *lamp*, and *chair* [37].

At the same time, when traveling up the object hierarchy, the levels become more abstract, and it becomes increasingly difficult to capture those levels with concrete representations. Back to the *fruit* example: the parent trees could consist of *food* and *drink*, for which any concrete representation would be inherently distanced from the abstract concept. Thus, the basic level is alternatively defined as "the most abstract level at which it is possible to have a relatively concrete image" [37].

Basic Levels in WordNet

The WordNet graph can be viewed as composing such an object hierarchy. Thus, when considering modifications to noun compounds through deviations based on WordNet synsets, it would not be surprising if the concept of the basic level came into play. In fact, the very existence of such a level makes a potential hypothesis like "moving up the WordNet graph makes a compound easier to interpret" problematic: depending on a compound's starting position in the object hierarchy vis-à-vis the basic level, deviations upward could leave it *too* abstract, just as deviations downward could leave it *too* specific.

Taking the above into consideration, a reasonable hypothesis would be that deviations towards the basic level lead to compounds that are more easily interpretable. Other reasonable hypotheses could be presented as well. Unfortunately, none of these are testable in practice. The idea of a basic level is mostly abstract and has not been extended to cover WordNet. Thus, as of now, it is not possible to determine the basic level, which makes it infeasible to assess any hypotheses that revolve around it.

As such, we present a simpler hypothesis that will likely fail to capture the nuance of the peer generation process: that movements down the WordNet graph, which increase specificity and, as a result, limit the creativity available to interpreters, make compounds more difficult to interpret. Note that in the basic level setting, we would expect this to be true some, but not all of the time.

Semantic Distance

Earlier, we demonstrated that the semantic distance between generated compounds and their attested variants correlated with difficulty of interpretation. By extending this logic, we would expect peer compounds that deviated more significantly from their root compound to have labels that differed more significantly from the original.

For example, say we have a compound `X Y`, and can swap `X` with the modifier `A` that is three synsets away or the modifier `B` that is just two synsets away. Under this reasoning, we would expect the label of `A Y` to differ more than `B Y` from that of `X Y`, given that `A` is further away from the original modifier `X`.

In the case of the WordNet relations available at present, then, we can explicitly compare the deviations of `Nephew` peers to those of `Child` peers, as the former will be definitively further away than the latter from the root compound (i.e., `Nephew` peers will always be three edges away in WordNet, while `Child` peers will be just a single hop away).

7.2 Hypotheses

The hypotheses in this section are simple and aim to capture the core motivations presented in Section 7.1 above:

H1. Deviations that move upward in the WordNet graph should more frequently *increase* ease of interpretability than those that move downward in the graph.

H2. Deviations that move a larger distance in WordNet should lead to larger deviations in interpretability. Specifically, `Nephew` peers should exhibit larger deviations than `Child` peers.

H3. Deviations enforced on a noun compound's modifier should be more substantial than deviations on a compound's head. By 'substantial', we do not mean that these deviations should necessarily make the compound easier or more difficult to interpret; rather, that they should have a more easily observable effect. This is because changing a compound's head can dramatically change its overall meaning, whereas changing a modifier will often leave the compound's core entity unchanged. For example, given the compound *fruit fly*, *apple fly* is closer in conception than, say, *fruit bug*, since the core entity ('fly') is retained.

Each of these hypotheses was considered as we conducted our analysis, and each will be addressed in the discussion below.

7.3 Data Generation

Next, we outline the method by which our initial set of 250 generated compounds were *mutated*, or altered using predictable deviations based on WordNet synsets, to produce a set of peer compounds.

To start, we defined four different WordNet relations that would be used when generating peers from an initial WordNet synset. These relations are listed in Table 1 and outlined visually in Figure 2 on Page 14. For completeness, they are: `Child`, `Sibling`, `Nephew`, and `Uncle`. Note that the `Uncle` relation goes upward, while the `Child` and `Nephew` relations go downward, and the `Sibling` relation stays on the same level of the WordNet graph.

For each noun compound from our initial round of experiments, we collected three distinct judgments from Turkers. Based on the paraphrases provided by Turkers, for each judgment, we were able to label the head and modifier of a compound with a specific WordNet synset. As a result, a compound's head and/or modifier could have been labeled with multiple different WordNet synsets across the three judgments submitted by Turkers; this multiplicity of synsets would be indicative of the multiple different senses in which the head and/or modifier was used during interpretation.

Thus, for each compound, for both its head and modifier, for every sense in which they were used, we generated every possible peer compound by iterating over the four pre-defined WordNet relations and producing a new peer for each `Child`, `Sibling`, `Nephew`, and `Uncle`. Specifically, for each synset collected as above, we iterated over its lemmas, as defined in WordNet, discarding any lemmas that were proper nouns, acronyms, hyphenated, or exhibited another unhelpful characteristic. Note that this method of data generation led to some noun compounds having multiple peers for a given relation, and others having as few as zero.

For example, in the initial dataset, we included the generated compound *country sugar*. Multiple Turkers provided paraphrases that led to *country* being assigned the synset *country.n.4*, and *sugar*, the synset *sugar.n.1*. By taking the `Nephew` relation from the modifier *country*, we reached the synset *village.n.2*, which produced the peer compound *village sugar*–note the increased specificity (from *country* to *village*) as we travel downwards in the WordNet graph via the `Nephew` relation. Alternatively, by taking the `Uncle` relation from the head *sugar*, we reached the synset *spice.n.2*, which produced the peer compound *country spice*.

To provide a level of quality control, we only used as root compounds those in the initial dataset with a clear majority-voted interpretability label (i.e., we excluded any compounds for which each of the three judges assigned a different interpretability label) and, in addition, we only used synsets for judgments whose label corresponded to that of the majority-vote winner.[18]

Recall that 219 compounds had a clear majority-voted interpretability label. For each of those compounds, we attempted to generate one peer by mutating the head and one by mutating the modifier, balancing the distribution of mutations across the dataset such that an even number of peer compounds were based on each relation, to the extent possible.

In the end, this process produced 351 peer compounds, with their mutating relations split along the lines of the values listed in Table 13. The final list of 351 peer compounds generated by this process can be found in Section B.3 of the Appendix.

[18]For example, if a compound received two *No difficulty* judgments and one *Minor difficulty* judgment, we would only use the synsets corresponding to the *No difficulty* judgments when generating its peers.

Relation	N_{all}	N_{head}	N_{mod}
Nephew	90	48	42
Uncle	92	46	46
Sibling	85	43	42
Child	84	42	42

Table 13: The number of peer compounds included for each WordNet relation type, where N_{head} indicates the number of compounds for which the head was mutated, and N_{mod}, for which the modifier was mutated.

7.4 Experiment Statistics

We provide a brief overview of the relevant experiment statistics in a manner similar to those presented in Section 5.1.

This second round of experiments was conducted in the window from November 18, 2014 to November 29, 2014. As the total number of compounds increased from 250 in the initial experiment to over 350, we allowed the batch size to increase to 100 compounds per batch.

Due to an experimental error, some batches mixed HITs pertaining to valid peer compounds with others that were not included in the final dataset. This made it difficult to gauge the average time per judgment. However, based on the average time for batches in which only valid peer compounds were included, we estimated that Turkers spent an average of 33 seconds on each HIT, a value below that of the previous round of experiments, but still in-line with our pre-experiment estimate.

Again, due to the aforementioned experimental error, the acceptance rate of submissions was difficult to compute exactly, but a similar estimate left us with an approximate acceptance rate of 93.88%, a number very close to the previous round's 93.75%.

Turker 'Diversity'

As we produced 351 peer compounds and collected three distinct judgments per compound, this made for 1,053 accepted HITs. These 1,053 HITs were submitted by a total of 99 different Turkers. The average number of accepted submissions per Turker was 10.64, while the median was 3. Again, there was a long tail of Turkers that submitted just a few HITs (73.3% of Turkers submitted between 1 and 10 HITs), and while no Turker hit the 50-HIT cap, two Turkers submitted exactly 49. These numbers are similar to those of the previous round of experiments.

Breakdown of Submissions

Of the 1,053 approved HITs, which spanned 351 distinct compounds, 415 submissions (39.4%) labeled a compound to be interpretable with *No difficulty*, 365 submissions (34.7%) labeled a compound interpretable with *Minor difficulty*, and 273 submissions (25.9%) labeled a compound *Meaningless*. Compared to the breakdown of submissions from the first round of

experiments provided in Section 5.2, these percentages differ by a raw margin of less than 2% for each interpretability label, a remarkably consistent result.

Majority-vote and unanimity breakdowns are also presented in Table 14. The majority-vote breakdown is very similar to that of the previous round of experiments (results differ by at most an absolute margin of 2.4%), although the unanimity percentages are slightly lower across-the-board.

Given that, for example, 38.4% of compounds received a majority-voted interpretability label of *No difficulty* in the first round of experiments and 36.8% were deemed as such in this second round, the consistency between experiments is indeed surprising. In fact, this consistency could suggest that the values presented in Table 14 are reasonable estimates for compound interpretability *in general*, not just in the context of this experiment.

Difficulty	**Num. Judgments**	**Num. Majority**	**Num. Unanimous**
No difficulty	415 (39.4%)	129 (36.8%)	50 (14.2%)
Minor difficulty	365 (34.7%)	104 (29.6%)	16 (4.56%)
Meaningless	273 (25.9%)	73 (20.8%)	23 (6.55%)

Table 14: Initial results from the second round of Amazon Mechanical Turk experiments, which consisted of 1,053 approved HITs spanning 351 distinct noun compounds, all of which were 'peers' of the compounds used in the first round of experiments.

7.5 Analysis

In this section, we examine how the paraphrases submitted by Turkers (and the synsets to which they were linked) matched up with those assigned to peer compounds through the peer generation process, with the explicit goal of addressing the hypotheses listed in Section 7.2.

The Assumption of Predictability

A key assumption underlying the hypotheses from Section 7.2 was that WordNet similarity would be an effective proxy for predicting how a user would interpret a given compound. This is distinct from predicting the difficulty of interpretation.

To be specific, when we generated these peer compounds, we did so by taking a synset (i.e., a sense) in which either the head or modifier of a compound had been interpreted and finding a replacement synset nearby in the WordNet graph, using this new synset to form a peer. Implicitly, then, this process assumed that when Turkers were presented with this peer compound, they would be inclined to interpret the new head or modifier in the sense represented by the WordNet synset used to generate it.

As will be clear from the analysis below, this assumption may have been invalid. In particular, we found that predicted peer synsets were used in only a minority of judgments. The implication is that Turkers were more likely to use some other synset corresponding to

the peer compound's new component than that which was predicted by the data generation process.

The issue at hand is non-obvious, but can hopefully be illuminated with an example. In our initial dataset, we included the generated compound *body manager*. Turkers generally interpreted the head, *manager*, with the synset *coach.n.1*, defined as "someone in charge of training an athlete or a team". When generating peers, one such peer, *body conditioner*, was based on the synset *conditioner.n.2*, defined as "a trainer of athletes" and connected to *coach.n.1* by the `Child` relation. However, when Turkers were presented with *body conditioner*, in all three judgments, *conditioner* was interpreted with the synset *conditioner.n.3*, defined as "a substance used in washing (clothing or hair) to make things softer". While we expected Turkers to interpret *conditioner* as some sort of trainer, they instead tended interpreting it as a grooming product.

To put it in simple terms: while we assumed that the peers we generated would be interpreted on the basis of the synsets from which they were produced, this was often not the case. This phenomenon is complicating, yet interesting in its own right. Specifically, since our goal was to evaluate how interpretability labels altered as compounds were mutated in *known* ways via *known* WordNet relations, such an evaluation cannot be undertaken for those compounds in which the predicted peer synset was not used by human judges. This limits the investigation of any hypotheses based on *known* WordNet relations to the portion of the dataset for which Turkers landed on synsets identified beforehand. However, it also opens the door for another portion of analysis, that which focuses on how frequently Turkers landed on the synsets identified beforehand as a function of the predicted relations. For example, were peers generated by the `Child` relation more frequently in sync with the predicted synsets than, say, peers generated by the `Uncle` relation? Or, alternatively: how often did an unpredictable interpretation involve a concept from a higher level in the WordNet graph, given that these higher levels house vaguer concepts?

In the sections below, we begin by discussing how the interpretability labels changed for those peers interpreted as predicted beforehand. Afterwards, we analyze the dataset through another lens, investigating the rate at which peers were judged in senses that deviated from those predicted by WordNet.

Changes in Interpretability

To start, we took the raw peer compound dataset (consisting of three distinct human judgments per peer) and annotated the head and modifier of each compound, for each judgment, with a WordNet synset, using the process described in Section 6.1. This required over 1,500 manual annotations, but was considered necessary to maintain quality.

Next, the dataset was segmented into those judgments for which the annotated synset matched the predicted synset decided upon during peer generation. Returning to the *body manager* and *body conditioner* example from above, any judgments that interpreted *conditioner* with the synset *conditioner.n.2* would be accepted, while those that interpreted it with any other synset would be discarded. In addition, only judgments with an interpretability label that agreed with the peer's majority-voted label were included, to mirror

the filtering process we enforced during the peer generation process from Section 7.3.

For each remaining judgment, for each WordNet relation, and for peers that were generated by both deviations from the head and modifier, we computed the number of judgments for which the human-provided interpretability label of a peer matched that of its root compound (e.g., for judgments using the synset *conditioner.n.2*, how frequently did *body conditioner*'s interpretability label match that of *body manager*?). We also counted the number of judgments for which the interpretability label was indicative of increased or decreased difficulty, respectively, as measured by movements from *No difficulty* to *Minor difficulty* and so forth.

Recall that, in the peer generation process, we used the first-sense heuristic to provide synsets for *Meaningless* compounds. Thus, deviations in interpretability labels could jump (in terms of becoming less difficult) from *Meaningless* to *No difficulty*; however, measurable improvements in difficulty could only go as high as increasing to *Minor difficulty*, since it was not possible for compounds deemed *Meaningless* to yield interpretations that matched their predicted WordNet synsets, given that these compounds were by definition not interpretable. To account for this discrepancy, we also report the number of peer compounds that ended up with a majority vote of *Meaningless*.

The results, for peers derived from deviating heads and modifiers respectively, are presented in Tables 15 and 16. In those tables, each cell contains a count of the number of peers created from a given WordNet relation for which the interpretability label differed (or not) from the original compound from which it was generated. The rows are divided into three sets of rows: the top set of rows tracks the number of peers that preserved the interpretability labels of their root compounds; the middle set, peers that became more difficult to interpret; and the bottom set, peers that became easier to interpret.

When parsing the results in Tables 15 and 16, recall that we included between 42 and 48 peers for every pair of WordNet relation and variant type (i.e., head or modifier). For example, there were 42 peers generated by taking the `Nephew` mutation from a compound's modifier. In addition, note that the modifications that end in *Meaningless* judgments are slightly inflated and should not be categorized in the same way as modifications ending in *No difficulty* or *Minor difficulty* labels, as a compound labeled *Meaningless* did not yield judgments using the predicted peer synset (as the judgments provided did not indicate *any* synset); the figures are merely included for completeness.

While there are inferences to be drawn from Tables 15 and 16, the most important observation is that there is a clear lack of data. Due to the discrepancy between predicted synsets and those used by human judges, along with the wide variety of WordNet relations that were explored, we have but a handful of eligible judgments for each permutation of settings (typically, around 10 peers per pair of WordNet relation and head or modifier variant). This makes it very, very difficult to draw firm conclusions and, in particular, to make definitive conclusions on the hypotheses we outlined in Section 7.2. On this basis, we defer most of the substantive analysis to the next section, where we turn this data sparsity into a strength by dissecting the rates at which predicted synsets matched those used by human judges.

However, there are still some observations worth noting. For one, it was incredibly rare for

	Child	Nephew	Uncle	Sibling
Stayed at *None*	6	5	5	3
Stayed at *Minor*	3	1	3	5
Stayed at *Meaningless*	1	1	2	1
None → *Minor*	4	2	1	4
Minor → *Meaningless*	4	5	3	2
None → *Meaningless*	2	4	4	5
Meaningless → *Minor*	0	0	0	0
Minor → *None*	2	2	2	4
Meaningless → *None*	0	0	0	0

Table 15: For peers based on mutations of the **head**, we tracked the rate at which interpretability labels changed, with respect to the labels of the root compounds from which they were generated. Each cell represents the number of peers that were given the resulting interpretability label (determined by majority vote) conditioned on the root compound's interpretability label (again determined by majority vote).

	Child	Nephew	Uncle	Sibling
Stayed at *None*	4	5	5	5
Stayed at *Minor*	1	3	1	7
Stayed at *Meaningless*	0	2	1	3
None → *Minor*	3	3	4	5
Minor → *Meaningless*	6	5	4	5
None → *Meaningless*	2	3	2	7
Meaningless → *Minor*	1	0	0	0
Minor → *None*	2	4	0	1
Meaningless → *None*	0	1	2	0

Table 16: For peers based on mutations of the **modifier**, we tracked the rate at which interpretability labels changed, with respect to the labels of the root compounds from which they were generated.

peers generated from *Meaningless* compounds to yield interpretations that fit those predicted at time of generation. This could speak to the inaccuracy of the first-sense heuristic (we have little proof that the first sense, which was used to generate the peer, accurately reflects a human judge's 'best guess' interpretation) or to the Power Law-esque phenomenon we have seen in the past: while it was common for compounds to go from *Minor difficulty* to *No difficulty*, it was uncommon for compounds to jump from *Meaningless* to some other gradient of 'interpretable'. However, in the absence of additional information, the noise induced by the first-sense heuristic provides a more acceptable explanation.

In general, there were no obvious differences between those peers generated by head and modifier manipulation, although this lack of difference likely speaks more to the sparsity of the dataset than anything else. For peers generated from both head and modifier manipulation of a *No difficulty* compound, it was relatively common to receive a *No difficulty* label; at least, slightly more common than those that stayed at a *Minor difficulty* label. In both cases, it was slightly more common for a (compound, peer) pair to go from *No difficulty* to *Minor difficulty* than vice versa.

Given these observations, we find ourselves unable to draw definitive conclusions as to the validity of the hypotheses presented in Section 7.2. While `Nephew` peers occasionally exhibited greater deviation in interpretability than `Child` peers, there's not enough data to differentiate signal from noise. Similarly, while `Child` and `Nephew` peers became easier to interpret at a higher rate than `Uncle` peers, the deviations are on the order of single digits. Further, the similarities between the values reported in Tables 15 and 16 make it difficult to differentiate between the effects induced by deviating heads versus those induced by deviating modifiers.

In this section, then, we established that WordNet deviations do *not* account for the productivity of noun compounds in a straightforward manner. Next, we analyze this finding in more detail by evaluating the degree to which productivity *was* captured by the peers generated using each of the four WordNet relationships considered during their creation.

Changes in Synset Usage

As a corollary to the lack of data present in the previous section, there were a large number of observations for which the synset corresponding to a peer judgment did not match up to the synset predicted at time of generation. Next, we explore how these *out-of-sync* judgments varied across the predicted WordNet relations.

The process by which relevant statistics were computed for these out-of-sync judgments ran as follows: For every pair of peer compound and root compound from which it was generated, we identified the synset used to produce the peer, as well as the predicted synset for that peer. Returning to the *body manager* and *body conditioner* example, the relevant synsets would be *coach.n.1* and *conditioner.n.2*. The former (*coach.n.1*) is included as it is the synset from *body manager* that was used to produce the peer, while the latter (*conditioner.n.2*) is included as it is the peer's predicted synset.

Next, we then filtered out any pair for which the root synset (e.g., *coach.n.1*) was used in just one of the three judgments submitted by Turkers. We also filtered out any pairs for which the peer had only one possible synset in WordNet, since this synset would be selected by default given the first-sense heuristic and would thus bias our results towards those peers for which there was only one available synset. In this way, the computed values, if anything, underrepresent the rate at which predictions matched up with human judgments.

After this filtration step, we computed:

- The number of times (out of three judgments) that the target peer synset was used in the interpretation provided by human judges.

- Whether the target peer synset was the most commonly occurring synset.
- Whether the target peer synset occurred *at all* in the three judgments.

Note that, as opposed to in the previous section, we made no effort to filter based on the interpretability labels of the judgments provided and instead restricted our view to the synsets.

As a concrete example, consider the potential root compound *fruit fly*, and assume that we received three judgments that were then best labeled with the modifier synsets *fruit.n.1* (indicative of an organic, editable fruit), *fruit.n.3* (indicative of the consequence of some action, as in "the fruit of one's labor"), and *fruit.n.1* again. We could then use *fruit.n.1* to generate the peer compound *orange fly*, where *orange* was derived from the predicted synset *orange.n.1* (indicative of orange the fruit), which is close to *fruit.n.1* in the WordNet graph.

When presenting human judges with this new compound, however, we might find that their judgments are best represented with the synsets *orange.n.2* (indicative of orange the color, rather than the fruit), *orange.n.2* again, and finally *orange.n.1*. In this case, we would include the pair of *fruit fly* and *orange fly* in our analysis, as the root synset, *fruit.n.1*, was included in two of the three human judgments. The most commonly-occurring synset in the peer judgments, *orange.n.2*, occurred twice, and the predicted synset, *orange.n.1*, occurred once. Thus, the final values for our three metrics, as computed on the pair of *fruit fly* and *orange fly*, would be 1, `False`, and `True`, respectively.

The results derived from this process, for peers generated by deviating heads and modifiers are presented in Tables 17 and 18, respectively.

	`Child`	`Nephew`	`Uncle`	`Sibling`
Mean predicted synset occurrences	0.82	*0.72*	0.89	**1.12**
Predicted synset most common (%)	29.63	28.12	34.62	**36.00**
Predicted synset never occurred (%)	66.67	*68.75*	61.54	**48.00**

Table 17: For peers based on mutations of the *head*, we report the rates at which synsets, as determined by user-submitted paraphrases, were in and out of sync with those predicted by WordNet. For each row, particularly large values are presented in **bold**, while particularly small values are presented in *italics*.

	`Child`	`Nephew`	`Uncle`	`Sibling`
Mean predicted synset occurrences	*0.50*	0.89	0.77	**0.80**
Predicted synset most common (%)	*20.83*	28.57	27.27	**33.33**
Predicted synset never occurred (%)	*79.19*	57.14	63.64	**53.33**

Table 18: For peers based on mutations of the *modifier*, we report the rates at which synsets, as determined by user-submitted paraphrases, were in and out of sync with those predicted by WordNet.

These tables paint a much clearer picture of the effect of WordNet relations on predictability of synsets. In particular, we make a few key observations:

- For peers generated by mutations of both the head and modifier, `Sibling` peers were much more likely to have been interpreted in a way that aligned with the synsets predicted during the data generation process. For example, the predicted synset was the most frequently-occurring synset for 36.00% and 33.33% of the `Sibling` peers that made it through the filtration step.

- Again for both mutations, `Child` and `Nephew` peers were, in general, the least in sync with predicted synsets. For example, nearly 80% of the `Child` synsets were *never* interpreted with the predicted synset for those peers generated by mutating a modifier.

- Prediction rates were generally similar between the peers generated by deviating heads and those generated by mutating modifiers, although predictions were *slightly* more accurate for those in the former category.

We can color these observations with a few examples. Namely, when looking at the second observation, it is interesting to see how the synsets derived from user judgments deviated from those predicted by WordNet relations. Often, these deviations arose due to the multiple possible senses in which the new peer component could be interpreted. For example, our initial dataset included the compound *player industry*. Turkers often interpreted *player* in the sporting sense of the word–as in, a *football player*. This gave us the `Child` peer *seed industry*, where *seed* was linked to the synset *seeded_player.n.1*, defined as: "one of the outstanding players in a tournament". But when this compound was presented to Turkers, they universally interpreted seed with the synset *seed.n.1*, defined as: "a small hard fruit". In this case, then, we have a peer compound (*seed industry*) for which the root (*player industry*) was easily interpretable with a specific sense, and the peer, easily interpretable with a very different sense. While our methodology expected *seed* to be interpreted in the sense of a *seeded* participant, Turkers instead saw it as the *seed* of a fruit. The *body manager* and *body conditioner* example from above is a similar case.

What these examples demonstrate is that, while WordNet distance and semantic similarity metrics may be useful when evaluating the interpretability of a compound relative to attested variants, they're not effective in predicting the precise way in which a compound will be interpreted, especially vis-à-vis generated compounds. This is a testament to the incredible productivity of noun compounds and the English language more generally: while the idea of a 'seeded player' is, obviously, closely linked to the concept of a 'player', the easiest way for humans to interpret the *seed industry* compound was to completely traverse the WordNet graph and come up with a completely different meaning of the word *seed*, which embodied a completely different semantic relationship to the head (*industry*).

In some cases, the divergence was a byproduct of WordNet's over- or under-specificity. We occasionally generated peers for which the predicted synsets were ultra-specific (like *quad.n.3*, defined as: "a block of type without a raised letter; used for spacing between words or sentences"). Indeed, it would be rare for human judges to tend toward these synsets when interpreting the compounds, which made this a difficult task from the start.

7.6 Conclusion

In this section, we generated a set of 250 peer compounds, which were constructed by mutating noun compounds from our initial dataset through preset WordNet relations. For each of these compounds, we collected three human judgments on their interpretability using the same HIT format as in Section 4.3.

In analyzing the results of our experiments, we found that WordNet proximity was an ineffective means of predicting the way in which a given compound would be interpreted. While this was not the hypotheses that we set out to validate, it is an interesting observation nonetheless and truly speaks to the complexity of noun compounds. That deviating compounds by following well-defined paths in the WordNet graph could lead to interpretations completely unrelated to those of the origin synsets, yet *still* mutually agreeable to human judges, is a fascinating discovery.

However, the conclusion is more nuanced than "WordNet proximity is a bad predictor for the interpretation of the peer". Rather, we found that it is difficult to decipher *when* peer relationships will be important, and how differences will emerge when mutating noun compounds. Going back to the *body manager* to *body conditioner* example: it is true that these peers appeared to be close in WordNet yet ended up with very different interpretations. But it also holds true that we can find peers of *body conditioner* with very *similar* interpretations. The peer *hair conditioner*, for example, would be a peer of *body conditioner*, reached by mutating the modifier *body*, with (in all likelihood) the same interpretation for *conditioner*.

In other words, then, the challenge is to discern which peers will be grouped together in terms of similarity of interpretation. For every compound, there are many peers for which the deviated component would be interpreted in a sense similar to that of the original; but it is untrue that *every* such peer would fall under this category.

Predicting that two peers will be clustered (i.e., use the same interpretations for their shared head or modifier) is not straightforward. In the context of the aforementioned example, the goal would be to develop a theory as to why are the peers *body manager* and *body conditioner* are dissimilar (in that *manager* and *conditioner* are not interpreted in the senses corresponding to close WordNet synsets), yet the peers *body conditioner* and *hair conditioner* are similar (in that *body* and *hair* would presumedly be interpreted in the senses corresponding to close WordNet synsets). Predicting peer clusters is beyond the scope of this study, but if anything, the results of this experiment demonstrate that the task is both difficult and fascinating.

8 Extending to Ternary Compounds

While noun compound research typically limits its scope to binary compounds, we choose to extend our study to ternary compounds as well.

When discussing questions of interpretability, ternary compounds are particularly interesting in that they themselves contain sub-compounds. Recall, for example, that the ternary

compound *olive oil bottle*, parsed as *[[olive oil] bottle]*, contains the sub-compound *olive oil*.[19] Our study aimed to identify how the interpretability of sub-compounds, like *olive oil*, could impact the interpretability of larger compounds of which they are a part, like *olive oil bottle*.

After completing the experiments outlined in the previous sections, we had access to 500 annotated compounds. This presented a unique opportunity to generate ternary compounds based on these annotated binary compounds. Given that we had a source of truth for the interpretability of the sub-compounds (human judgments from the HITs), we were able to analyze, for the first time, how the interpretability of a sub-compound impacts a larger compound's overall interpretability.

Again, we used the AMT platform to collect data on the interpretability of ternary compounds. Following the HIT template described in Section 8.2, we gathered 1,800 human judgments, with the results presented and analyzed in Section 8.5.

8.1 Hypotheses

Given a new ternary compound of the form `X Y Z`, there are two possible ways to bracket it: `[[X Y] Z]` and `[X [Y Z]]`. In other words, one can interpret `X Y Z` as a left- or right-branching compound, respectively, based on the sub-compound that is preserved when defining its interpretation.

This multiplicity raises interesting questions about the interpretability of ternary compounds. For one, it is often possible to produce two very different interpretations for the same compound based on the direction of branching. The compound *club cover charge*, for example, could be interpreted as "the cover charge required to attend a night club", or, just as easily, "the cost of a golf club cover". These two interpretations correspond to bracketing the compound as right- or left-branching, respectively. As such, there's some sense in which ternary compounds allow for *increased* creativity or optionality during parsing, since human judges can explore both left- and right-branching interpretations.

Note that, in generating our data, we produced ternary compounds `X Y Z` such that one of `X Y` and `Y Z` was an attested compound, and the other, a generated compound with a known interpretability label. Thus, each ternary compound considered can be viewed as a mashup of two binary compounds, one of which a human judge would likely be familiar, and the other, unfamiliar. For example, by combining the generated compound *horse war* and the attested compound *war criminal*, we produced the ternary compound *horse war criminal*. The exact procedure by which these ternary compounds were generated is described in more detail in Section 8.3 below.

Given the format of the data and the assumption of increased creativity, we developed the following hypotheses:

[19] The bracketing syntax used here was introduced in Section 2.1 on Page 8. In brief, when a ternary compound `X Y Z` is bracketed as `[[X Y] Z]`, the claim is that the interpretation corresponding to this bracketing involves a grouping of the words `X Y`. In the *olive oil bottle* example, bracketing the compound *[[olive oil] bottle* is in sync with the interpretation "a bottle of olive oil", as the sub-compound *olive oil* remains intact in this definition.

H1. Human judges lean towards branching on a familiar compound. In particular, Turkers will branch more often in the direction of the attested compound, such that they provide a left-branching paraphrase for the compound `X Y Z` if `X Y` is attested and a right-branching paraphrase if `Y Z` is attested.

H2. Human judges are able to interpret ternary compounds containing *Meaningless* generated sub-compounds, more often than not, due to the familiarity of the attested sub-compound and the direction it provides in parsing a larger ternary compound. In other words, when combining a *Meaningless* compound and an attested compound to produce a ternary compound, the presence of the attested compound will aid human judges as they develop valid interpretations.

H3. There will be a larger proportion of *Minor difficulty* ternary compounds than there were binary compounds. This is partly motivated by the expanded set of potential interpretations for ternary compounds, (made possible by the branching effect described above, which allows for increased creativity and optionality during parsing) and partly motivated by the increased entropy of ternary compounds (a function of the extra word vis-à-vis binary compounds, which increases the chance that any two words in the compound will pair together awkwardly).

H4. Ternary compounds will often share the interpretability label of the generated compound that they contain. The labels will overlap more frequently when the judge has chosen to branch along the path of the generated compound (i.e., when the judge has chosen to keep the generated compound together in the provided paraphrase, they will be more likely to have provided the same interpretability label than if they had kept the attested compound together).

As in the previous rounds of experiments, our HIT template was designed to collect data that would validate or invalidate these hypotheses.

8.2 Human Intelligence Task Format

We designed our HITs so as to collect sufficient data to validate or invalidate the hypotheses listed in Section 8.1. In particular, for each ternary compound, we wanted to collect:

- An interpretability label on the scale of *No difficulty*, *Minor difficulty*, and *Meaningless*.
- A paraphrase that captured the Turker's mental model for interpreting the compound. Most importantly, this paraphrase had to demonstrate whether the Turker was interpreting the compound as left- or right-branching.

This second point is of particular importance. In the initial round of experiments, we were collecting paraphrases to gain a better sense for how Turkers were interpreting the compounds put before them. These paraphrases allowed us to assign WordNet synsets to the head and modifier of each compound, detect the diversity of interpretations, and so forth.

In this round of experiments, we simplified our template to focus on whether the compound was left- or right-branching, and less on the Turker's specific interpretation. This was a design decision that made the experiment easier to follow for subjects, while retaining all the necessary information for assessing the hypotheses listed in Section 8.1.

Determining the Direction of Branching

However, determining whether a compound was interpreted as left- or right-branching was an unprecedented challenge. Historically, there's been very little work on the task of determining the direction in which a compound should branch, and even *less* work on the task of determining these directions based on human-provided paraphrases. The closest work is from [21], in which the authors use algorithmic techniques to decide on whether a compound should be considered left- or right-branching.

On the AMT platform, these determinations often involve more art than science. In effect, the goal was to trick Turkers into revealing the manner in which they were mentally branching during interpretation–without explicitly introducing the concept of branching, as the mere mention of the idea would render the task overly difficult. As such, we had to develop a clever system for collecting branching decisions. This system had to be both (1) simple enough for Turkers to understand, and (2) accurate enough for us to infer whether a compound was left- or right-branching, based on input from a human judge.

In the end, our HIT template ran as follows: first, each HIT asked the Turker to decide on an interpretability label for the ternary compound in a manner similar to that of the previous round of experiments. Next, the Turker was asked to provide a paraphrase for the compound. This paraphrase could take *any* format, with one restriction: assuming that the ternary compound took the form `X Y Z`, the paraphrase was required to preserve either `X Y` or `Y Z`. Later, the compound was deemed left-branching if the Turker preserved `X Y`, and right-branching if the Turker preserved `Y Z`.

Returning to the *club cover charge* example, if a Turker submitted the paraphrase "the cover charge required to attend a club", we would deem the compound right-branching, as *cover charge* was preserved in the interpretation. On the other hand, given the paraphrase "the charge for a club cover", we would deem the compound left-branching, as the paraphrase preserved the sub-compound *club cover*. In particular, we would say that the latter paraphrase *branched with* or *preserved* the sub-compound *club cover*. These two phrases will be used interchangeably below.

While this technique is based on a heuristic, we found it to be quite accurate, especially given that it required minimal explanation or understanding on the part of human judges.

It should be noted, however, that this system left us with paraphrases that were:

1. More difficult to parse programmatically, as they no longer followed a known format, as opposed to those collected in the previous rounds of experiments, which were based on the relative clause or use of a preposition, as described in Section 4.2.

2. Less explicit. For example, a Turker could paraphrase a ternary compound of the form `X Y Z` as "a `Y Z` for `X`". This would provide us with the compound's branching designation,

but as the Turker was not required to provide a paraphrase for the sub-compound `Y Z`, the exact interpretation of the overall compound was not always entirely clear.[20]

In evaluating these tradeoffs, we opted for a simpler experiment that more directly addressed our hypotheses.

HIT submissions were rejected based on the criteria as in Section 4.3. Submissions were also rejected if they failed to keep `X Y` or `Y Z` together in the paraphrase provided.

A sample HIT can be found in Section C.2 in the Appendix.

8.3 Data Generation

Recall that ternary compounds take the form `X Y Z` for three distinct nouns `X`, `Y`, and `Z`. In our experiments, we wanted to make claims about how the interpretability of a ternary compound `X Y Z` is impacted by the interpretability of its sub-compounds, `X Y` and `Y Z`.

As such, we generated our ternary compounds using the following procedure:

1. Pick a generated binary compound from the previous round of experiments (i.e., a compound with a known interpretability label). Only binary compounds with a clearly majority-voted interpretability label were used, i.e., those for which we had at least two of the three judgments in agreement over the choice of interpretability label.

2. Decide whether the generated compound should compose the left or right two words of the overall ternary compound. For example, assuming the final ternary compound took the form `X Y Z`, we could use the generated compound as either the sub-compound `X Y`, thus positioning it on the left, or the sub-compound `Y Z`, positioning it on the right.

3. If the generated compound was positioned on the left of the ternary compound, then pick an attested compound of the form `Y Z` and merge the generated compound `X Y` with the attested compound `Y Z` to create the ternary compound `X Y Z`. Otherwise, pick an attested compound of the form `X Y` and continue as above.

The goal was to merge a generated and attested compound by finding a pair of compounds with a shared head and modifier, or shared modifier and head. For example, the annotated compound *candy eye* could be merged with the attested compound *eye movement* to create the left-rooted compound *candy eye movement*. Alternatively, *candy eye* would be merged with the attested compound *sugar candy* to create the right-rooted compound *sugar candy eye*.

8.4 Experimental Design

The procedure described in Section 8.3 is parameterized on two arguments: the interpretability label of the generated binary sub-compound, which could be either *No difficulty*, *Minor*

[20] While we toyed with the idea of requiring a second paraphrase for the sub-compound preserved by the Turker, this would have complicated the experiment and distracted us from our primary goals.

difficulty, or *Meaningless*, as defined in Section 2.4); and the position of the generated sub-compound within the larger ternary compound (left or right). This leads to six distinct combinations of parameters.

As our hypotheses in Section 8.1 concerned a variety of different parameter combinations, we choose to generate 100 ternary compounds for each combination, making for 600 total ternary compounds.

As in the first round of experiments (Section 4), we collected three judgments per compound, making for 1,800 total HITs. Each HIT followed the template outlined in Section 8.2. Again, as in the experiments of Section 4, Turkers were limited to at most 100 submissions each given the size of the dataset. The HITs were batched in groups of 100, also an increase over the batching size of Section 4, and again due to the larger the dataset.

8.5 Results

We begin with a high-level overview of the experiments statistics in a manner similar to that of Section 5.1. That is, we focus here on the number of participants, a breakdown of the interpretability labels provided, and so forth. More thorough, conclusion-driven analysis will be presented in Section 8.6 below.

The experiments were conducted over a seven day window from February 26, 2015 to March 4, 2015. In collecting 1,800 total judgments (three for each of the 60 ternary compounds), HITs were batched in groups of 100, making for 18 batches in total.

The acceptance rate for submissions was 81.06%, which represents a 12% drop from the acceptance rates of the previous experiments (i.e., those conducted over binary compounds). This lower acceptance rate was quite clearly a function of the increased complexity of the task. In many cases, Turkers evidently failed to read the instructions closely and provided paraphrases that merely consisted of two of the three words in the compound (e.g., submitting just the words "fruit fly" to paraphrase the ternary compound *orange fruit fly*), or example sentences containing the compound; in both of these scenarios, the HITs were rejected. As in the previous rounds of experiments, collecting paraphrases from human judges allowed for a degree of quality control.

Turker 'Diversity'

In collecting 1,800 HITs, we approved submissions from 116 different Turkers. On average, 15.52 HITs were accepted per Turker; the median number of accepted HITs was 7.

As in the previous experiments, there was a long tail of Turkers with a small number of submissions, and a small group of Turkers who hit or came very close to the per-Turker limit of 50 accepted HITs. Specifically, 56.9% of Turkers submitted fewer than 10 accepted HITs, and 13 Turkers (11.2%) submitted between 45 and 50, with 8 (6.90%) hitting the cap of 50.

Breakdown of Submissions

Our experiment included 600 distinct ternary compounds. As in the previous rounds of experiments, we collected three judgments per compound, making for 1,800 total accepted HITs. Of these 1,800 HITs, 692 submissions (38.4%) labeled a compound to be interpretable with *No difficulty*, 620 submissions (34.4%) labeled a compound interpretable with *Minor difficulty*, and 488 submissions (27.1%) labeled a compound *Meaningless*. These figures are presented in Table 19. The percentages differ by a raw margin of less than 3% when compared to the breakdown of submissions from the first round of binary experiments (Section 5.2), and a margin of 2% when compared to the submissions from the second round of binary experiments (Section 7.4). Again, the rates at which judges attribute these labels are remarkably consistent.

Majority-vote and unanimity breakdowns are also presented in Table 19. The majority-vote breakdown is very similar to that of the first round of binary compound experiments, differing by at most 2% in any of the three labels. However, the unanimity percentages are lower across-the-board. It is interesting to note that the pattern of lowest-unanimity for *Minor difficulty* compounds, which we saw in Section 5.2, persists in Table 19.

Difficulty	**Num. Judgments**	**Num. Majority**	**Num. Unanimous**
No difficulty	692 (38.4%)	221 (36.8%)	75 (12.5%)
Minor difficulty	620 (34.4%)	170 (28.3%)	24 (4.00%)
Meaningless	488 (27.1%)	138 (23.0%)	41 (6.83%)

Table 19: Initial results from the Amazon Mechanical Turk experiments on ternary compounds, which consisted of 1,800 approved HITs spanning 600 distinct noun compounds.

In relating the results of Table 19, and, in particular, the figures in that table vis-à-vis those of the binary experiments from Section 5.2, to the hypotheses from Section 8.1, we can see that the proportion of *Minor difficulty* judgments did *not* increase when expanding to ternary compounds, contrary to **H3**. Between these two sets of experiments, the overall percentage of *Minor difficulty* judgments dropped by 1.6%, although the percentage of majority- and unanimously-voted *Minor difficulty* compounds increased by 1.1% and 1.0% respectively. However, overall, these deviations are surprisingly small, and the rates at which interpretability labels are assigned seems consistent across the binary and ternary compound-based experiments.

8.6 Analysis

Next, we analyze the results of our experiments in greater detail. Of particular interest to us (the experimenters) is the way in which Turkers chose to branch in their interpretations of the ternary compounds and, specifically, the degree to which their choice of branch correlated with the position of the attested sub-compound and/or the interpretability label of the generated sub-compound, as determined in the first round of experiments.

Branching Directions

Each ternary compound `X Y Z` implicitly contains two binary sub-compounds: `X Y` and `Y Z`. When generating the ternary compounds used in our experiments, exactly one of `X Y` and `Y Z` would be a generated binary compound from our first round of binary experiments, constructed according to the process described in Section 4.1; the other, an attested compound from one of the existing binary datasets, as listed in Section 2.4.

In Section 8.1, and in **H1** in particular, we claimed that human judges would tend to branch along the path of the attested sub-compound, as keeping that sub-compound together would make the overall ternary compound easier to interpret, given the increased familiarity of the attested sub-compound vis-à-vis the generated sub-compound.

As an example, consider the ternary compound *soft drink room*, which consists of the attested compound *soft drink* and the generated compound *drink room*. In this case, it is more intuitive for a human judge to interpret the compound as *[[soft drink] room]* (e.g., "a room in which one consumes soft drinks") than as *[soft [drink room]]* (e.g., "a drink room that is soft"), given that *soft drink* is a familiar, commonly-occurring binary compound. By bracketing the compound as *[[soft drink] room]*, the judge would be branching in the direction of the attested sub-compound, rather than the generated.

With this established, we now assess the validity of hypothesis **H1** from Section 8.1. To start, we note that of the 1,800 total judgments received from Turkers, 1,312 of them used a *No difficulty* or *Minor difficulty* label. In these cases, judges were required to submit a paraphrase, which was used to determine the direction of branching (see Section 8.2 for more). Of those 1,312 judgments for which a compound's branch direction was determinable, 913 deemed the compound to be left-branching, and 399, right-branching. Alternatively put, 69.6% of judgments deemed a compound left-branching.[21] This value is roughly in line with the range of 64% to 67% proposed by Lauer [21], based on statistical analysis of a corpus of noun compounds.

However, when the attested sub-compound was on the left, human judges branched left in 568 judgments and right in just 101 judgments, making left-branching interpretations over five times as popular. Alternatively, when the attested sub-compound was on the right, human judges branched right in 298 and left in 345 judgments, an almost even split. These figures are presented in Table 20.

From Table 20, the effect of the attested sub-compound in anchoring interpretations is quite clear: whereas judges branched in the direction of a left-positioned attested sub-compound nearly 85% of the time, they preserved right-positioned attested sub-compounds only 53.7% of the time, a nearly even split. The difference between these two percentages is statistically significant at a 99% confidence level using a standard Z-test for proportions.

In conclusion, then, we consider **H1** to be valid in that human judges tended to preserve attested sub-compounds when interpreting unfamiliar ternary compounds. This was identified by the higher rates at which judges branched left or right depending on whether the attested sub-compound was positioned, similarly, on the left or right of the overall ternary

[21] When determining branch direction by a majority-vote, 73.7% of compounds were deemed left-branching.

Group	Left-Branching	Right-Branching
All	913 (69.6%)	399 (30.4%)
Left attested	568 (**84.9%**)	101 (15.1%)
Right attested	345 (53.7%)	298 (**46.3%**)

Table 20: The direction in which human judges branched when paraphrasing ternary compounds, with judgments segmented by the position of the attested sub-compound within the larger ternary compound. Note that, for example, "Left attested" indicates that the attested sub-compound was on the left. In general, human judges tended to preserve attested sub-compounds at disproportionate rates; certain exceptional values are **bolded** for clarity.

compound. In other words: the position of the attested sub-compound played a hugely influential role in determining the direction in which human judges tended to branch.

Branching as a Function of Interpretability Labels

We can further segment our judgments based on the interpretability labels provided by human judges. For each interpretability label (at least, for those with which a Turker was required to submit a paraphrase, namely *No difficulty* and *Minor difficulty*), we computed the same values as in the previous section–specifically, the breakdown between branching directions with respect to the position of the attested sub-compound. The results are presented Table 16.

	Group	Left-Branching	Right-Branching
No difficulty	All	504 (72.8%)	188 (27.2%)
	Left attested	321 (87.2%)	47 (12.8%)
	Right attested	183 (56.5%)	141 (43.5%)
Minor difficulty	All	409 (66.0%)	211 (34.0%)
	Left attested	247 (82.1%)	54 (17.9%)
	Right attested	162 (50.8%)	157 (49.2%)

Figure 16: Branching directions for ternary compounds segmented by interpretability label, position of the attested sub-compound, and direction of branching. Note that for *Meaningless* judgments, no such paraphrase was provided, which made the branching direction impossible to determine; hence, those judgments are omitted from consideration.

From Table 16, we can see that *No difficulty* judgments were more often those with a left-branching interpretation (72.8% vs. 66.0% for *Minor difficulty* judgments). Naturally, this increase carries over to the next two rows, which demonstrate that *No difficulty* judgments with a left-positioned attested sub-compound were more frequently left-branching than those for *Minor difficulty* judgments under the same conditions. The same observation can be made

for ternary compounds with right-positioned attested sub-compounds.

The discrepancy could speak to the relative ease of interpretation for left-branching compounds, given that they're slightly more natural and naturally occurring, according to the present study and that of Lauer [21]. Further, this strong preference for left-branching interpretations is consistent with the general principles of English syntax. As such, it could be the case that compounds which lend themselves to left-branching interpretations are generally easier to interpret given that such interpretations are more natural and more syntactically aligned with English syntax. Such a result would imply that a compound's branching designation could act as an indicator of its regularity and, further, its interpretability.

Interpretability Overlap

We next evaluate the degree to which the interpretability label of a ternary compound mirrored that of the generated sub-compound that it contained. Recall that for each of the generated sub-compounds, we collected three human judgments using the HIT format described in Section 4.2. Each judgment included an interpretability label, either *No difficulty*, *Minor difficulty*, or *Meaningless*. Given these three judgments, we could then assign a 'ground truth' interpretability label to the judgment by taking the majority vote (or acknowledge that the compound received three different interpretability labels and thus had no clear majority).

When generating ternary compounds, we exclusively used the generated binary sub-compounds with a clear majority-voted interpretability label from the last round of experiments, as described in Section 8.3. Thus, when comparing the label of a ternary compound to that of the generated sub-compound it contains, we always had a 'ground truth' interpretability label for that latter. For the ternary compound itself, we again took a majority vote over the judgments submitted by Turkers and discarded those compounds for which a clear majority did not exist.

Under these qualifications, we can then compare the interpretability labels of ternary compounds to those of their generated sub-compounds. The exact results are presented in Table 21. Most importantly, we found that ternary compounds agreed with the interpretability labels of their generated sub-compounds 45.6% of the time. This 45.6% label agreement rate is astoundingly high given the three-tiered approach to labeling. That figure in particular seems to support or even validate hypothesis **H4** from Section 8.1, which claimed that this agreement rate would indeed be non-negligible.

Table 21 contains four other rows. The 'Left-attested' and 'Right-attested' rows segment compounds based on the position of the attested sub-compound within the larger ternary compound. With these figures, we see that the position of the attested sub-compound played a minimal role in encouraging agreement or disagreement with the generated sub-compound.

The next two rows, those labeled 'Preserved attested' and 'Preserved generated', segment the ternary compounds based on whether the branching direction (as determined by a majority vote) went along the path of the attested or generated sub-compound.[22] In the *soft drink*

[22]Note that for these two rows, we could not include any compounds for which the generated sub-compound

Group	Agreement	Disagreement
All	241 (45.6%)	288 (54.4%)
Left-attested	120 (44.9%)	147 (55.1%)
Right-attested	121 (46.2%)	141 (53.8%)
Preserved attested	106 (40.6%)	155 (59.4%)
Preserved generated	59 (53.6%)	51 (46.4%)

Table 21: The rates at which the interpretability labels of ternary compounds (determined by majority vote) agreed with those of the generated binary sub-compounds they contained. In this table, ternary compounds are segmented based on the position of their attested sub-compound (rows 2 and 3), and then whether judges kept the attested or generated sub-compounds together in their interpretations (rows 4 and 5).

room example: if Turkers had deemed this compound left-branching (keeping the attested sub-compound *soft drink* together), it would be included in the 'Preserved attested' category; otherwise, if they had deemed it right-branching (keeping the generated sub-compound *drink room* together), it would be included in the 'Preserved generated' category.

These last two rows of Table 21 are of interest to us in evaluating the hypotheses of Section 8.1, and **H4** in particular, which claimed that interpretability labels would overlap more frequently when judges had branched along the path of the generated sub-compound, rather than the attested sub-compound. And, indeed, we see that interpretability labels were in agreement with those of their generated sub-compounds 53.6% of the time when branching with said generated sub-compounds versus just 40.6% of the time when branching with the attested sub-compound. Again, **H4** is validated by the results of Table 21.

In this section, then, we gained a better understanding of the effect that a sub-compound could play on a larger ternary compound of which it is a part. In particular, we saw a remarkably high agreement rate (45.6%) between the labels of generated sub-compounds and those of their larger ternary compounds. In addition, when segmenting based on the branching direction determined by Turkers, interpretability label agreement rates were significantly higher when Turkers branched in the direction such that kept the generated sub-compound was preserved in their paraphrase.

The Effect of Meaningless Sub-Compounds

In the hypotheses listed in Section 8.1, we included **H2**, which made the claim that ternary compounds with *Meaningless* generated sub-compounds would often be interpretable, despite the difficulty of interpretation of its component. For example, the ternary compound *state dinner officer* is composed of the attested sub-compound *state dinner* and the generated sub-

had a *Meaningless* interpretability label. This is because, if the ternary compound also received a *Meaningless* interpretability label, we would not be able to verify that judges had preserved the generated sub-compound, since at most one paraphrase would have been provided. Analysis of *Meaningless* generated sub-compounds is reserved for the next section, below.

compound *dinner officer*. The compound *dinner officer* was deemed *Meaningless* by judges in the first round of experiments. However, *state dinner officer* was deemed interpretable with *Minor difficulty* in this round of experiments, as the judges unanimously branched left, keeping *state dinner* together in their paraphrases. In a sense, the added context and information in the ternary compound made it much easier to interpret.

In the last section, when observing the effect that branching played on the agreement of interpretability labels for ternary compounds and their sub-compounds, we had to ignore those ternary compounds based on generated sub-compounds with *Meaningless* labels. Now, we exclusively look at the effect that *Meaningless* sub-compounds played on interpretability.

Of the 171 ternary compounds with both a clear majority-voted interpretability label and a *Meaningless* generated sub-compound, only 68 (39.77%) of them were deemed *Meaningless* by judges, with 42 (24.56%) deemed interpretable with *No difficulty* and 61 (35.67%) deemed interpretable with *Minor difficulty*. In other words: in accordance with **H4**, more often than not, compounds with a *Meaningless* sub-compound were nonetheless interpretable (although typically with some difficulty given the prevalence of *Minor difficulty* judgments).

Of the 103 ternary compounds with a *Meaningless* root but an overall *No difficulty* or *Minor difficulty* majority-voted interpretability label, 84 branched in the direction of the attested compound. Thus, 81.6% of these compounds eschewed an interpretation based on the generated sub-compound. This is the scenario described in our *state dinner officer* example, where judges were unable to interpret *dinner officer* in our first round of experiments, but were able to interpret this larger ternary compound by relying on the attested sub-compound *state dinner*.

In comparison to ternary compounds based on *No difficulty* and *Minor difficulty* sub-compounds, this proportion is highly inflated. In fact, between the various rates at which judges preserved generated sub-compounds, there's a clear trend: when the generated sub-compound was interpretable with *No difficulty*, judges preserved it 42.5% of the time; when interpretable with *Minor difficulty*, they preserved it just 33.9% of the time; and, taking the complement of the above, when *Meaningless*, they preserved it 22.8% of the time. The conclusion: *when sub-compounds are difficult to interpret, judges shy away from them; when they're easy, they're more likely to be preserved.*

These figures demonstrate both the flexibility of ternary compounds (i.e., that judges could eschew sub-compounds that were difficult to interpret in favor of more reasonable mental groupings, namely by keeping attested sub-compounds together) and the effect that sub-compounds play in enforcing certain interpretations. In evaluating our hypotheses, it is clear that with larger compounds come greater freedom and creative license in interpretation, as Turkers were able to interpret compounds containing *Meaningless* sub-compounds with relative ease. Future work could examine the nature of interpretability for even *larger* compounds (say, length-4 or length-5). In those cases, the problem becomes even more interesting, as larger compounds become awkward and difficult to manage, creating a tension between the vast realm of possible groupings and interpretations, and the unfamiliarity and awkwardness of these larger structures.

8.7 Conclusion

In this section, we devised and analyzed in experiment based on *ternary* compounds, or those consisting of three separate nouns, like *olive oil bottle*. While previous research has mostly been restricted to the analysis of binary compounds, we choose to examine the interpretability of ternary compounds, with a particular focus on the effect that the interpretability of binary sub-compounds could play on that of larger ternary compounds, of which they are a part.

Our experiment, which was conducted using the AMT platform, collected 1,800 distinct human judgments spanning 600 ternary compounds.

Before analyzing our results, we first presented a series of hypotheses in Section 8.1, which were then evaluated throughout Sections 8.5 and 8.6.

In considering the hypotheses of Section 8.1, we came to a number of conclusions regarding the interpretability of ternary compounds, many of which were in line with our intuition about human behavior. For example: when a ternary compound contained both a sub-compound that was very difficult to interpret and one that was not as challenging, judges tended to preserve the latter during parsing. Furthermore, in general, judges tended to preserve attested sub-compounds, regardless of whether they composed the two leftmost or rightmost words in the ternary compound. In many cases, this inertial effect allowed ternary compounds containing meaningless or nonsensical sub-compounds to themselves be open to valid interpretations, as judges could split apart those meaningless sub-compounds and fallback to preserving the attested sub-compounds with which they were likely familiar.

These conclusions speak to a theory of interpretation for ternary compounds in which the sub-compounds that compose them play an exceptionally important role. Additionally, this role appears to be non-linear in the sense that combining two compounds that are somewhat difficult to interpret does not appear to be a more productive method of generating interpretable ternary compounds than, say, combining one compound that is very easy to interpret and another that is very difficult.

9 Discussion

In this thesis, we set out to develop a theory of noun compound interpretability. While noun compound research has grown in scale and scope over the past few years, our efforts focused on the rarely asked questions of: "What makes a noun compound interpretable?" and, more broadly, "What are the limits of noun compound interpretability and productivity?" Given the innovative aspects of our experiments and the nature of our results, this thesis represents both a novel effort and a series of original contributions to the field of computational linguistics.

In the preceding pages, we dissected the results of three separate rounds of experiments, each of which was conducted on Amazon's Mechanical Turk platform. Each of these experiments was designed so as to include some degree of noun compound interpretation by human judges (Turkers), typically in reference to noun compounds that were assumed to be

unfamiliar or unusual.

At this juncture, we attempt to tie together the results of these three rounds of experiments–which were discussed in Sections 6, 7, and 8, respectively–with the goal of forming a consistent set of conclusions relating to the interpretability of noun compounds.

Rates of Interpretability

In **H4** of Section 3, we hypothesized that *most* noun compounds would be interpretable given the productivity and generativity of compounds as a linguistic structure. And indeed, in each round of experiments, we found most noun compounds to be interpretable, be it with ease or some degree of difficult. Given that the compounds in our dataset were constructed in a manner so as to make them as unusual as possible, this is a fascinating result, and one that speaks to the productivity and diversity of noun compounds: even with high entropy compounds, Turkers were generally able to form reasonable interpretations, as exemplified by the paraphrases they composed. Furthermore, our experiments demonstrated not only a high rate of interpretability, but also a surprising level of co-agreement among judges (see Section 5.1). In other words, not only were most compounds interpretable, but judges were generally able to agree on the degree of difficulty involved in their interpretation.

As a related result, we found that in each round of experiments, the set of compounds involved was judged to be interpretable at nearly identical rates. In other words, whether they were generated, peer, or ternary compounds, the proportions of compounds judged to be easily interpretable, interpretable with difficulty, and uninterpretable were very much aligned across datasets. Given the manner in which our datasets were constructed, the consistency of these interpretability rates suggests that they could in fact represent very general proportions. That is, these proportions could be capturing, to some degree, the rates at which *all* nouns compounds are interpretable, an observation that extends beyond those included in our datasets.

In summary, not only did we find that most compounds were interpretable, but also, that judges were generally able to agree on a compound's interpretability. Further, the proportions of compounds deemed to be easily interpretable, interpretable with difficulty, and uninterpretable were remarkably consistent across multiple experiments and datasets, suggesting that there may be a natural split between compounds in general.

The Interdependence of Paraphrasing and Interpretation

In addition to collecting raw interpretability labels, we also required judges to submit a paraphrase for each compound following a specific format based on use of prepositions or the relative clause. In **H5** of Section 3, we hypothesized that paraphrases corresponding to compounds deemed more difficult to interpret would be more complex, as evidenced by a greater diversity of tokens and other quantifiable metrics, with the broader implication being that the complexity of a paraphrase could be indicative of the difficulty involved in its composition.

Throughout this thesis, we have used these paraphrases to augment our analysis. Their

usefulness has led us to conclude that the acts of paraphrasing and interpretation are *intrinsically linked*: in effect, paraphrases reveal information about the difficulty of their composition, which in turn reveals information about the difficulty associated with interpreting a given compound. In other words, a compound's difficulty of interpretation and the complexity of its paraphrases are *not* independent.

This connection was evidenced on multiple occasions and for multiple definitions of 'complexity', extending beyond the characterization based on token diversity introduced in **H5**.

First, in Section 6.2, where we demonstrated that difficulty of interpretation correlated with the length of a paraphrase and the diversity of its tokens, a shallow definition of complexity that corresponded to our initial formulation from Section 3.

Second, in Section 6.5, where the topics and clusters computed over structural representations of paraphrases were linked to difficulty of interpretation, representing a more advanced definition of complexity.

And third, in Section 6.6, where vector-space representations of paraphrase dependencies were used to train a machine learning classifier to predict a compound's interpretability, an arguably deeper take on complexity. In each case, the paraphrases associated with a given compound were linked to that compound's interpretability, be it on a macro or micro scale, and using the actual content of the paraphrases or deeper structural representations.

The existence of a link between interpretation and paraphrasing is very much in tune with our intuition. However, to demonstrate the existence of this link experimentally, and on so many levels, from content- to structure-based approaches, is not only a novel contribution, but one that certainly merits further exploration.

A Comparison-Based Model of Interpretation

Perhaps the most significant analysis in this thesis focused on the usefulness of drawing comparisons between compounds that had been produced algorithmically (with which judges were assumed to be unfamiliar) and attested compounds found in existing noun compound datasets (with which judges were assumed to be familiar).

These comparisons relied on measures of semantic and lexical similarity, especially those based on WordNet [10], and were typically drawn between a generated compound (like *cotton cup*) and the attested compounds with which it shared a common term (like *cotton farmer*, which shares the modifier *cotton*, or *coffee cup*, which shares the head *cup*), known as *attested variants*.

In **H3** of Section 3, we hypothesized that these comparisons would be useful in gauging noun compound interpretability given (1) the application of semantic similarity measures in prior research and (2) the assumed role that comparisons to familiar compounds could play in interpreting new ones. And, indeed, throughout our analysis, we found these comparisons to be powerful and effective indicators of interpretability. In fact, these comparisons were typically more reliable than any of the other techniques explored.

We point to two areas in which these comparisons were used extensively and effectively: first, in demonstrating a correlation between semantic similarity (of a generated compound and its attested variants) and difficulty of interpretation (Section 6.3); and second, in the

training of a machine learning classifier using pairwise compound comparisons as its unit of account, with feature vectors computed using a variety of semantic and lexical similarity metrics (Section 6.6).

In both of cases, comparisons to attested compounds were identified as a crucial indicator of compound interpretability. One might be tempted to claim, then, that compounds proven to be more semantically similar to their attested variants are easier to interpret. However, the conclusion is not that simple. And, in fact, our analysis leads to a more nuanced observation: that comparisons to a *select group* of variants are more effective. In other words, while comparisons to a wide range of attested variants can be useful, identifying the variants of maximal importance can yield better results. For example, in training our machine learning classifier, we found that clustering attested variants, identifying the cluster of maximal similarity, and subsequently excluding any variants outside of that cluster improved performance substantially. This process was akin to cutting out irrelevant comparisons and, instead, limiting ourselves to a more relevant group of attested variants. Similarly, in assessing the correlation between semantic similarity and interpretability, we found that trends were most pronounced when we employed a 'best vector' approach, i.e., discarded every attested variant besides that which was maximally similar.

These two examples demonstrate that, while broad comparisons to a large pool of attested compounds can be useful, it is in fact better to identify the most relevant subset of attested variants and compare to this subset instead. Typically, the noun compounds in which a given word might be used can be classified or segmented based on senses in which that word is used and the semantic relationships in which it may be involved.

For example, compounds based on the pattern (∗ *cup*) could be divided into those that use *cup* as a liquid container and those that use *cup* as a trophy, and then into those that include a modifier representing a liquid, or a material, or whatnot. When considering the interpretability of a new, unfamiliar compound, our results seem to demonstrate the value of identifying those attested compounds that employ the sense or semantic relationship that is most similar to those employed in this unfamiliar compound, rather than comparing to *every* possible usage of its constituent components. Returning to the *cup* example: when developing a valid interpretation for *cotton cup*, it may be most useful to identify the subset of (∗ *cup*) compounds that use *cup* in a similar sense, rather than draw on every such compound in existence.

For further evidence of the importance of identifying a 'most relevant' subset of attested variants, we look to the experiments conducted on peer compounds (Section 7). We found that, while we had assumed that peer compounds would be judged in a sense predicted beforehand based on WordNet distances, instead, judges tended to move away from our predicted senses.

Recall that our peer compounds are themselves (unattested) variants. For example, the generated compound *player industry* produced the peer *seed industry*, where *seed* was meant to reference the idea of a seeded player in a tournament; in reality, however, Turkers interpreted *seed* as the seed of a fruit or vegetable. The difficulty, then, was that we were unable to predict peers' interpretability labels because we simply could not identify whether

a peer was *truly* a peer–in the sense of lending itself to a similar interpretation. In some cases, we generated variants (like *seed industry*) that were interpreted in ways that were worlds apart from that of their root compound (in this case, *player industry*). We failed, in other words, to identify whether two peers could be reasonably grouped together in the same cluster of variants, which rendered many of our hypotheses untestable. In effect, we had generated peers that went beyond the boundaries of a set of similarly interpretable compounds. The power of these variant clusters had a huge bearing on our results.

Composing Comparisons

Stepping back, it is important to note that the very usefulness of drawing comparisons between generated and attested compounds could be seen as evidence in favor of a compositional approach to noun compound interpretability, similar to that expressed by the Principle of Compositionality introduced in Section 2.4.

In particular, one might argue, based on our results, for a theory of interpretation in which a human judge, when presented with an unfamiliar compound, first identifies the set of possible senses in which its constituent components are used in familiar compounds, and then finds a mutually agreeable combination of senses for the two words to form a reasonable interpretation.

For example, in interpreting *cotton cup*, a judge could rely on drawing comparisons to the compounds *cotton shirt*–in which *cotton* is used as a material of which the head, *shirt*, is composed–and *paper cup*–in which *cup* is used as a container, constructed out of the material referenced in the modifier, *paper*–to interpret the compound as a "cup made of cotton".

Under this model, the interpretation of a noun compound would involve not only the composition of the meanings associated with its member terms, but rather, of the ways in which these member terms are used in the wild (i.e., in attested variants). This could be seen as an extension of the Principle of Compositionality in which the candidate set of senses and semantic relationships considered for a given word are guided by inferences and comparisons to familiar variants, a theory that is simple, compelling, and evidenced, to an extent.

Beyond the Binary

Beyond our analysis of binary compounds, we ran an experiment to explore the interpretability of *ternary* compounds, with a particular focus on the degree to which the interpretability of a sub-compound affects that of a larger parent compound (Section 8). Given that compound research has traditionally focused on binary compounds exclusively, we feel that our use of ternary compounds represents another significant contribution to the field, especially given the burden of developing innovative forms of data collection, for which there was no precedent, as described in Section 8.2.

In **H2** of Section 8.1, we proposed that some ternary noun compounds containing *Meaningless* binary sub-compounds would nonetheless be interpretable. In validating this hypothesis, our analysis demonstrated that the interpretability of larger noun compounds, and ternary compounds in particular, is an entirely non-linear phenomenon. For example,

a ternary compound composed of two sub-compounds that are themselves considered to be relatively easy to interpret could, in fact, prove more difficult to interpret than a compound composed of one very easily interpretable and one meaningless binary sub-compound. This phenomenon relies on the ability to 'branch' when interpreting a ternary compound. That is, when parsing a ternary compound of the form `X Y Z`, one can either preserve the sub-compound `X Y` or the sub-compound `Y Z`. The very existence of this choice allows for a greater degree of freedom and creates a dominating effect in which the maximal interpretability of any sub-compound is more important than, say, the average interpretability of all sub-compounds involved.

In generating our dataset, we formed each ternary compound by combining a familiar (*attested*) and unfamiliar (*generated*) binary compound, as described in Section 8.3. Given this construction, we hypothesized in **H1** of Section 8.1 that in interpreting ternary compounds, judges would generally preserve the *attested* sub-compounds with which they were familiar, and in **H4**, that when preserving a *generated* sub-compound, they would likely provide the overall ternary compound with an interpretability similar or identical to that of the generated sub-compound.

In our analysis, we found both of these hypotheses to be validated, as human judges tended to preserve attested sub-compounds when interpreting larger ternary compounds. In the event that a judge instead preserved a generated sub-compound, they were more likely to label the ternary compound in agreement with the interpretability label of that generated sub-compound. Interestingly, generated sub-compounds were preserved more frequently when they were easy to interpret or, at the very least, interpretable with some difficulty (based on the interpretability labels provided in the initial round of experiments on binary compounds).

These observations can be seen as evidence of an *anchoring* phenomenon in the interpretation of ternary compounds: in effect, judges tended towards the sub-compounds with which they were presumedly most familiar, preferring simpler interpretations to those that demand elaborate explanation, a result reminiscent of Occam's Razor. Even when Turkers preserved an unfamiliar generated sub-compound, it was often the case that this sub-compound had been deemed easy to interpret or interpretable with some difficult. And although we would not be able to prove such a claim without further vetting of the attested sub-compounds, it may be the case that some of these situations involved generated sub-compounds were more familiar to human judges than the attested sub-compounds.

In conclusion, our analysis of ternary compounds suggests that compound interpretation is a non-linear process, not only in terms of ease of interpretation, but also in the manner in which interpretations are constructed. For example, based on the above, we could posit that developing an interpretation for `X Y Z` is more of a question of "Which of `X Y` and `Y Z` is easier to interpret" than, say, "Which of `X Y` and `Y Z` is easier to interpret *given* `Z` and `X`". Simply put, the interpretability of a ternary compound is more a function of the maximal interpretability of its sub-compounds than anything else. This claim itself is interesting in that it very much views ternary compound interpretation as a function of sub-compounds, rather than individual words, a view that we hold to be well-evidenced in its correctness.

Towards a More Comprehensive Understanding of Interpretability

In this section, we tied together the various pieces of analysis introduced throughout this thesis, with the goal of developing a number of conclusions as to the interpretability of noun compounds. The theory we have presented above speaks to the diversity and productivity of compounds as a linguistic structure, but does not portray them as a construct beyond our comprehension.

In particular, compounds appear to be intimately linked to those that share common terms, and comparisons between a given compound–especially one that is considered unfamiliar or unusual–and its attested variants can go a long way towards determining its interpretability. The fact that these comparisons were useful is itself evidence as to the importance of semantic and lexical similarity in noun compound interpretation, given that similarity-based metrics laid the foundation upon which these comparisons were drawn.

Further, the very interpretation of compounds was found to be intrinsically linked to the act of paraphrasing, as paraphrases frequently leaked information as to the difficulty involved in their composition. Just as comparisons to attested variants were useful in evaluating a compound's interpretability, so too was the manner in which it was paraphrased by human judges. In theory, then, a compound's paraphrases could act as a proxy for its interpretability label.

The conclusions scattered throughout this thesis extend beyond these closing observations and are almost certainly connected in non-obvious ways. Perhaps this connective tissue will be illuminated in the future as we continue to improve on our understanding of noun compound interpretability. But for now, to summarize compounds with a single thesis statement would be to fly in the face of all that make them great.

10 Conclusion

In this thesis, we explored the question of what makes a noun compound interpretable. Through a series of experiments conducted on Amazon's Mechanical Turk platform, we collected human judgments on the interpretability of a set of algorithmically-generated binary and ternary compounds, and used this original dataset to test a series of hypotheses in a scientifically rigorous manner.

Our results, as discussed in Section 9, demonstrate that the interpretation of unfamiliar noun compounds is an act that relies on drawing comparisons to familiar, attested compounds. This process can be modeled using measures of semantic and lexical similarity, particularly those based on WordNet [10]. Further, we showed that the acts of interpretation and paraphrasing are intrinsically linked, and that the interpretation of larger compounds (ternary compounds, in particular) lends itself to a non-linear formulation in which the maximal interpretability of any sub-compound is a key factor.

However, beyond these conclusions, each of which represents an original contribution to the field of computational linguistics and linguistics more generally, this thesis contained several other innovations, mostly related to the algorithmic construction of noun compounds

as well as the scientific collection and analysis of interpretability judgments. The peer construction process outlined in Section 7.3 and the ternary branch-designation scheme devised in Section 8.2 exemplify these contributions well.

While our results are promising and led us to a number of satisfying and interesting conclusions, this thesis paved the way for a great deal of future work. In particular, running similar experiments on a much larger dataset (think: thousands or even millions of compounds, rather than hundreds) would be wise, especially for training a machine learning classifier (Section 6.6). Luckily, the data generation process introduced in Section 4.1 is sufficiently simple, allowing for the painless construction of massive datasets, which could be used for future experimentation.

Additionally, in Section 9, we raised the issue of identifying the most relevant cluster of peer compounds for a given generated compound, highlighting its importance. While our analysis in Section 6.6 did make use of unsupervised graph clustering algorithms to determine a set of maximally relevant peers, our approach was relatively untested and very few alternatives were considered. As such, this task deserves further consideration.

Our exploration of ternary compounds (Section 8) was itself a novel endeavor, as much of the existing research on noun compounds has been restricted in scope to binary compounds. It would be of great value for researchers to carry out similar experiments over sets of larger compounds (say, length-4 or length-5). It is our belief that as compounds grow in size, they face a tradeoff between the increased space of possible interpretations (given the branching permutations available to judges) and the unwieldy nature of their composition (as massive compounds are often awkward). While ternary compounds in general merit further exploration, the question of compound interpretability as a function of size is one of particular interest to these authors.

Further, as discussed in Section 9, paraphrases and interpretability labels exhibited a surprisingly strong connection, which we have described as an *intrinsic link*. This result is particularly relevant to existing noun compound research. Much of the prior work on compounds has focused on the act of paraphrasing [27]. Developing more complex models through which to draw inferences as to the interpretability of a compound based on its paraphrases could prove useful to those academics focused on paraphrase generation. For example, in training an automated paraphrase generator, researchers may be able to improve performance in practice by ignoring those noun compounds that are not genuinely interpretable by human judges; in that case, a paraphrase complexity-based model could be used to discern the interpretable from the uninterpretable. While this is but one example, it nonetheless demonstrates the existence of a substantial set of unanswered questions with regards to noun compound interpretation and paraphrasing.

As such, our hope is that the results presented in this thesis will inspire researchers to re-examine the question of noun compound interpretability. While noun compounds research has come under the spotlight recently, the questions explored in this paper had been largely ignored before publication. And as demonstrated above, these questions not only merit exploration in their own right, but further, answering them can help us develop a more comprehensive understanding of noun compounds in general. Such a theory would

extend beyond the determination of a compound's interpretability; rather, it would have very general implications, even affecting the tasks on which researchers are already so focused, like the development of semantic classification taxonomies and the automatic generation of paraphrases.

References

[1] K. Ahn, J. Bos, D. Kor, M. Nissim, B. L. Webber, and J. R. Curran. Question Answering with QED at TREC 2005. In *TREC*, 2005.

[2] K. Barker and S. Szpakowicz. Semi-Automatic Recognition of Noun Modifier Relationships, 1998.

[3] L. Bauer and A. Renouf. A Corpus-Based Study of Compounding in English. *Journal of English Linguistics*, 29(2):101–123, 2001.

[4] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[7] M.-C. de Marneffe and C. D. Manning. Stanford Typed Dependencies Manual. Technical report, Stanford University, 2008.

[8] M.-C. de Marneffe and C. D. Manning. The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[9] N. C. Ellis. Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, 24(02):143–188, 2002.

[10] C. Fellbaum. WordNet: An Electronic Lexical Database, 1998.

[11] B. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[12] E.-H. Han and G. Karypis. Centroid-Based Document Classification: Analysis and Experimental Results. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 424–431, London, UK, UK, 2000. Springer-Verlag.

[13] K. M. Hermann, P. Blunsom, and S. Pulman. An Unsupervised Ranking Model for Noun-noun Compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 132–141, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[14] A. Huang. Similarity Measures for Text Document Clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, pages 49–56, 2008.

[15] P. G. Ipeirotis, F. Provost, and J. Wang. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA, 2010. ACM.

[16] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, Jul 2002.

[17] S. N. Kim and T. Baldwin. Automatic Interpretation of Noun Compounds Using WordNet Similarity. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP'05, pages 945–956, Berlin, Heidelberg, 2005. Springer-Verlag.

[18] D. Klein and C. D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[19] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[20] M. Lauer. Corpus Statistics Meet the Noun Compound: Some Empirical Results. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 47–54, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[21] M. Lauer. Designing Statistical Language Learners: Experiments on Noun Compounds. Technical report, 1995.

[22] D. Lee, H. Chuang, and K. Seamons. Document Ranking and the Vector-Space Model. *Software, IEEE*, 14(2):67–75, Mar 1997.

[23] J. N. Levi. The Syntax and Semantics of Complex Nominals, 1978.

[24] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[25] L. Meng, R. Huang, and J. Gu. A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1), 2013.

[26] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, 2011.

[27] P. I. Nakov and M. A. Hearst. Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases. *ACM Trans. Speech Lang. Process.*, 10(3):13:1–13:51, July 2013.

[28] V. Nastase and M. Hearst. Exploring Noun-Modifier Semantic Relations. In *Proceedings of the International Workshop on Computational Semantics*, IWCS '03, pages 285–301, 2003.

[29] F. J. Newmeyer. Review of 'The Syntax and Semantics of Complex Nominals'. *Language*, 55(2):pp. 396–407, 1979.

[30] E. Nicoladis. What's the Difference Between 'Toilet Paper' and 'Paper Toilet'? French-English Bilingual Children's Crosslinguistic Transfer in Compound Nouns. *Journal of Child Language*, 29(04):843–863, 2002.

[31] D. Ó Séaghdha and A. Copestake. Co-occurrence Contexts for Noun Compound Interpretation. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 57–64, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[32] D. Ó Séaghdha and A. Copestake. Using Lexical and Relational Similarity to Classify Semantic Relations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 621–629, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[33] B. H. Partee. Lexical Semantics and Compositionality. *An Invitation to Cognitive Science: Language*, 1:311–360, 1995.

[34] A. Peñas and E. Ovchinnikova. Unsupervised Acquisition of Axioms to Paraphrase Noun Compounds and Genitives. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'12, pages 388–401, Berlin, Heidelberg, 2012. Springer-Verlag.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] J. Ramos. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[37] E. Rosch, C. B. Mervis, W. D. Gray, D. M, and P. Boyes-Braem. Basic Objects in Natural Categories. *Cognitive Psychology*, 1976.

[38] V. Rus, M. Lintean, C. Moldovan, W. Baggett, N. Niraula, and B. Morgan. The SIMILAR Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts. In *Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May*, pages 23–25, 2012.

[39] V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu. SEMILAR: The Semantic Similarity Toolkit. In *ACL (Conference System Demonstrations)*, pages 163–168. Citeseer, 2013.

[40] H. S. Sichel. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, 70(351a):542–547, 1975.

[41] S. Tratz and E. Hovy. A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 678–687, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[42] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[43] L. Vanderwende. Algorithm for Automatic Interpretation of Noun Sequences. In *COLING '94 Proceedings of the 15th conference on Computational linguistics - Volume 2*. also available as no. MSR-TR-94-21, MSR techical report, ACL, 1994.

[44] S.-M. J. Wong, M. Dras, M. Johnson, et al. Topic Modeling for Native Language Identification. 2011.

A Source Code & Data

The relevant source code used to analyze the datasets gathered from Amazon Mechanical Turk is publicly available as a git repository hosted on the Bitbucket platform. Within that repo, much of the code of interest is contained in a series of IPython notebooks, which can be found in the `./src/analysis` directory.

The raw data collected from the AMT platform as well as the synset annotations (as described in Section 6.1) will be made available upon request.

B Experiments on Binary Compounds

B.1 Binary Dataset

The initial round of experiments on binary compounds consisted of collecting three judgments for each of a set of 250 compounds, which were generated as per the process outlined in Section 4.1. The exact compounds interpreted by human judges were as follows:

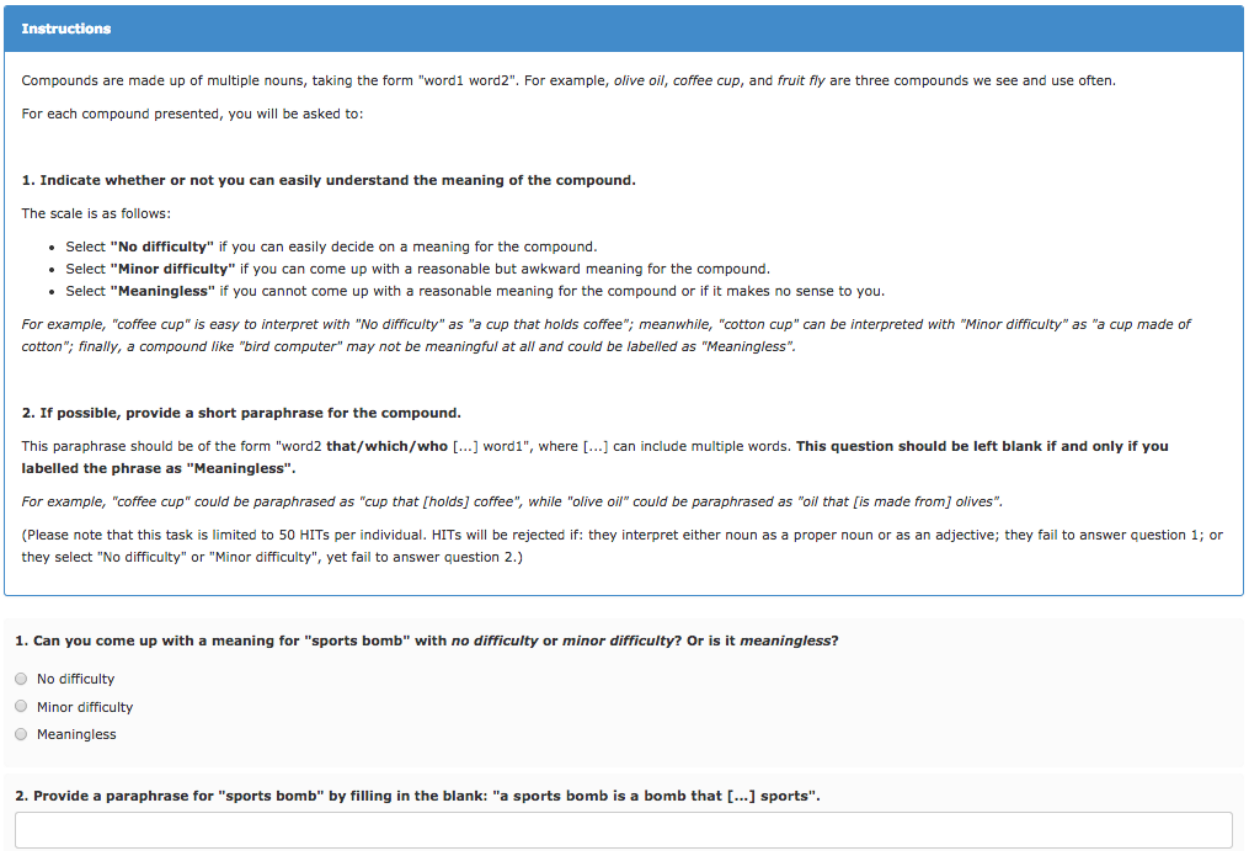
1. pressure dispute
2. government bars
3. bronze charge
4. pet members
5. bank function
6. string victim
7. account consultant
8. world policy
9. drink room
10. government eye
11. power analysis
12. wastebasket attempts
13. air zone
14. starvation shock
15. post field
16. country sugar
17. sugar measure
18. daisy baby
19. pork wall
20. decision development
21. bus function
22. novelty employees
23. concrete colonies
24. summer dispute
25. party cab
26. review identity
27. vacuum range
28. wax area
29. pole donor
30. future officer
31. citizen teams
32. peanut folk
33. machine actor
34. chocolate burn
35. paper attempts
36. desert hours
37. beer fair
38. bacon sauce
39. mining donor
40. birth disease
41. training arena
42. research expression
43. exhibition dish
44. ship members
45. sea lake
46. control office
47. pet limb
48. mountain utensils
49. rice eye
50. life zone
51. draft food
52. product step

53. job advance
54. sea machine
55. water book
56. chain concession
57. cigarette helmet
58. business party
59. nomination survey
60. care party
61. air lodge
62. daisy deaths
63. student pressure
64. golf purpose
65. top group
66. sports creations
67. city members
68. tax doctor
69. phantom nation
70. automobile dune
71. pet lock
72. city dispute
73. canine oil
74. surface period
75. extension friends
76. subject decision
77. accident dispute
78. house structure
79. beard alcohol
80. siege door
81. future actor
82. peer version
83. pet problems
84. neighborhood cube
85. farmer research
86. machine core
87. margin office
88. testtube height
89. career practice
90. enemy signals
91. dinner officer
92. draft plant
93. drug orders
94. acting fair
95. marriage lane
96. nut engineer
97. college committee
98. student paper
99. wastebasket inventions
100. plum deaths
101. college limb
102. castle decision
103. apple alcohol
104. steam soldier
105. iron area
106. cathedral performance
107. student pains
108. motor time
109. city engineer
110. home jar
111. horse war
112. government power
113. voice creations
114. coal supplies
115. policy limb
116. history requests
117. water control
118. jungle range
119. neighborhood lake
120. starvation problems
121. water cure
122. bull signals
123. pet cake
124. citizen activities
125. petrol addict
126. beehive machine
127. exercise disease
128. cathedral administration
129. hotel model
130. cigarette pains
131. bear helmet
132. retirement practice
133. blanket months
134. growth lakes
135. ground manager
136. body instrument
137. resource revenue
138. part subject

139. input jacket
140. domain army
141. waste application
142. assistance assembly
143. hand food
144. company glass
145. law attack
146. adventure invasion
147. faculty machine
148. wind administration
149. back number
150. handlebar assets
151. jungle paper
152. frog ring
153. interest man
154. air work
155. child church
156. charity case
157. heat height
158. desert prayers
159. suspense attack
160. dualist service
161. search speed
162. teaching friends
163. deficiency committee
164. country machine
165. salmon datum
166. moth force
167. operating official
168. home creations
169. heart members
170. surface colonies
171. community stock
172. terry arrangement
173. love colonies
174. player industry
175. soya office
176. body manager
177. enterprise product
178. door control
179. security arrangement
180. oil ring
181. part decision
182. headquarters protocol
183. cotton order
184. lightning measure
185. siphon letter
186. video example
187. exhibition competition
188. sports bomb
189. hermit committee
190. family analysis
191. lightning country
192. deficiency control
193. vapor item
194. child fish
195. computer work
196. assistance base
197. enemy cure
198. cigarette attack
199. trust money
200. government statue
201. horse structure
202. language inventions
203. discharge control
204. disaster bread
205. abbey assembly
206. mother butter
207. candy eye
208. sports price
209. factory lecture
210. college weather
211. hydrogen bomb
212. mother requests
213. nose money
214. language construction
215. hand charge
216. flu machine
217. organ area
218. flounder magazine
219. disease imports
220. picture procedure
221. jungle force
222. food firm
223. sugar bars
224. gardening tool

225. teaching church
226. job paper
227. engineering treaty
228. snow sugar
229. bath centre
230. aerospace project
231. garden study
232. accident burn
233. home song
234. government structure
235. citizen helmet
236. charity report
237. child limb
238. lounge provision
239. party soup
240. vegetable rash
241. fitness experience
242. extension magazine
243. faculty remedy
244. enemy prayers
245. picture shape
246. holiday manager
247. faculty lake
248. sap folk
249. fatigue country
250. loan approach

B.2 Sample HIT: Binary Compounds



Instructions

Compounds are made up of multiple nouns, taking the form "word1 word2". For example, *olive oil*, *coffee cup*, and *fruit fly* are three compounds we see and use often.

For each compound presented, you will be asked to:

1. Indicate whether or not you can easily understand the meaning of the compound.

The scale is as follows:

- Select **"No difficulty"** if you can easily decide on a meaning for the compound.
- Select **"Minor difficulty"** if you can come up with a reasonable but awkward meaning for the compound.
- Select **"Meaningless"** if you cannot come up with a reasonable meaning for the compound or if it makes no sense to you.

For example, "coffee cup" is easy to interpret with "No difficulty" as "a cup that holds coffee"; meanwhile, "cotton cup" can be interpreted with "Minor difficulty" as "a cup made of cotton"; finally, a compound like "bird computer" may not be meaningful at all and could be labelled as "Meaningless".

2. If possible, provide a short paraphrase for the compound.

This paraphrase should be of the form "word2 **that/which/who** [...] word1", where [...] can include multiple words. **This question should be left blank if and only if you labelled the phrase as "Meaningless".**

For example, "coffee cup" could be paraphrased as "cup that [holds] coffee", while "olive oil" could be paraphrased as "oil that [is made from] olives".

(Please note that this task is limited to 50 HITs per individual. HITs will be rejected if: they interpret either noun as a proper noun or as an adjective; they fail to answer question 1; or they select "No difficulty" or "Minor difficulty", yet fail to answer question 2.)

1. Can you come up with a meaning for "sports bomb" with *no difficulty* or *minor difficulty*? Or is it *meaningless*?

- ○ No difficulty
- ○ Minor difficulty
- ○ Meaningless

2. Provide a paraphrase for "sports bomb" by filling in the blank: "a sports bomb is a bomb that [...] sports".

B.3 Peer Dataset

The second round of experiments on binary compounds consisted of collecting three judgments for each of a set of 250 *peer* compounds, which were generated as per the process outlined in Section 7.3. Note that each peer compound will share either a head or a modifier with one of the binary compounds listed in Section B.1 above.

The exact peer compounds interpreted by human judges were as follows:

1. meteor manager
2. part call
3. business product
4. career last
5. enemy input
6. eye money
7. bear sailor
8. phantom league
9. lightning collection
10. grease area
11. school committee
12. culture tool
13. court bars
14. role assembly
15. oil necklace
16. spectator actor
17. settlement machine
18. fitness step
19. hotel simulation
20. video sample
21. court statue
22. government can
23. cycling creations
24. drink offset
25. palladium bomb
26. water prescription
27. giant party
28. food giant
29. child bird
30. privatisation shock
31. winery lecture
32. love gown
33. fair competition
34. accident infection
35. strategy limb
36. factory offer
37. auditor pressure
38. exhibition tournament
39. brush range
40. post campus
41. beehive press
42. neighborhood pond
43. flight remedy
44. drink firm
45. fiber application
46. rainforest force
47. surface relay
48. duty party
49. circuit prayers
50. air center
51. care interest
52. court power
53. string target
54. research traffic
55. sinking colonies
56. screen control
57. major paper
58. college branch
59. top world
60. party truck
61. sports valuation
62. interest author
63. garden credentials
64. country shovel
65. trust cash
66. world docket
67. buffer assets
68. testtube gauge
69. picture condition
70. farmer probe
71. floor colonies

72. bank report
73. plantation study
74. assistance garrison
75. moth grip
76. beer form
77. retirement misconduct
78. motor hour
79. record example
80. review sunshine
81. services practice
82. spirits fair
83. derivative control
84. father requests
85. drug word
86. jungle plot
87. set disease
88. cow signals
89. enemy inquiry
90. birth breakdown
91. size control
92. growth branch
93. government span
94. authority analysis
95. stick helmet
96. picture tower
97. child leg
98. payroll revenue
99. system signals
100. family correction
101. wind set
102. church structure
103. payroll arrangement
104. peer letter
105. outlet protocol
106. chip function
107. prosecution speed
108. blanket date
109. propane supplies
110. lounge stockpile
111. control spa
112. faculty tablet
113. fetus ring
114. horse separation
115. hermit force
116. greyhound problems
117. urchin fish
118. ownership survey
119. polarisation analysis
120. gardening bar
121. player cooperative
122. sect glass
123. period officer
124. summer division
125. poster war
126. margin headquarters
127. policy bill
128. date alcohol
129. rainforest paper
130. infant church
131. government term
132. marriage channel
133. rainfall administration
134. yesterday number
135. resource gain
136. wafer eye
137. ground executive
138. city controversy
139. peanut scheme
140. computer behavior
141. discharge development
142. system core
143. pork screen
144. breakup disease
145. production dish
146. base cure
147. greyhound cake
148. honey measure
149. jungle sheet
150. water interest
151. decision reversal
152. hand juice
153. cotton pass
154. city participant
155. veto development
156. pet position
157. shrine performance

158. dropout committee
159. signing construction
160. company consultant
161. freshwater book
162. stretch friends
163. charity estimate
164. folk machine
165. cathedral executive
166. fall burn
167. signing inventions
168. press time
169. handlebar deficit
170. faculty sorter
171. town engineer
172. photograph shape
173. assistance recording
174. adolescent limb
175. air organization
176. nomination scan
177. escape control
178. ginger folk
179. back benefit
180. defence structure
181. power study
182. chenille victim
183. credit office
184. technology friends
185. starvation breakdown
186. relation version
187. water album
188. dualist facility
189. realty money
190. waste habit
191. heart sister
192. body conditioner
193. high group
194. training court
195. interference official
196. pressure explanation
197. chocolate infection
198. tool donor
199. country salt
200. organ pit
201. sea surface
202. tomorrow actor
203. ship tribesman
204. nationalist teams
205. canine helmet
206. aluminium area
207. teaching mortuary
208. castle saving
209. coal increase
210. daisy boy
211. jute order
212. ownership base
213. stage lake
214. photography price
215. party puree
216. iron property
217. due man
218. city specialist
219. neck sauce
220. golf target
221. jet measure
222. hand thinner
223. butterfly force
224. strength office
225. company system
226. devil nation
227. extension relation
228. shortage committee
229. community part
230. sugar ornament
231. winter dispute
232. petroleum addict
233. seed industry
234. dog oil
235. art procedure
236. straw height
237. frog bracelet
238. wax stadium
239. citizen section
240. horse line
241. flagship members
242. enterprise clothing
243. relationship members

244. student disaster
245. poppy baby
246. child dormitory
247. freshness experience
248. company stock
249. canine fat
250. celebration cab
251. carbohydrate ring
252. bank commission
253. reserve dispute
254. bacon salt
255. operating prosecutor
256. search length
257. machine dancer
258. index food
259. rice contractor
260. siege fence
261. student hiatus
262. language pornography
263. aerospace regimen
264. voter activities
265. engineering insurance
266. court model
267. pet cookie
268. business funeral
269. defense door
270. nut programmer
271. account counselor
272. bank utensils
273. cover arena
274. mountain steel
275. starvation innocence
276. tank centre
277. package machine
278. petrol junkie
279. enterprise invasion
280. comfort months
281. deficiency model
282. deficiency panel
283. hydrogen release
284. cement colonies
285. pony structure
286. citizen migration
287. draft poison
288. thigh room
289. pole supporter
290. pass project
291. teaching repeater
292. exercise collapse
293. mother tale
294. language representation
295. environment prayers
296. vacuum zone
297. accident division
298. counter work
299. session soup
300. machine bull
301. future comic
302. bolt country
303. door regulator
304. civilian helmet
305. bread burn
306. signing creations
307. puncture dispute
308. tailpipe letter
309. exhibition boat
310. apple elixir
311. acting presentation
312. security measure
313. sugar people
314. bus commission
315. washing paper
316. house shelter
317. citizen sailor
318. grower research
319. duct instrument
320. preparation practice
321. seafood wall
322. jungle outfit
323. town members
324. right charge
325. job magazine
326. chain franchise
327. relation policy
328. cathedral exposure
329. space engineer

330. concrete outpost
331. part machine
332. round purpose
333. medication orders
334. government control
335. charity engagement
336. bay provision
337. communication fair
338. home layer
339. parasite service
340. religion case
341. brokerage function
342. future procurator
343. bath lodge
344. voice sum
345. lesson church
346. survey expression
347. soil manager
348. lightning resistance
349. environment sugar
350. biotechnology treaty
351. disaster jumble

C Experiments on Ternary Compounds

C.1 Ternary Dataset

The experiments on ternary compounds were carried out over a set of 600 compounds, generated as per the process outlined in Section 8.3. The exact compounds interpreted by human judges were as follows:

1. surface period face
2. disease imports furniture
3. apple alcohol poisoning
4. neighborhood lake area
5. advertising account consultant
6. dawn air zone
7. domain army spokesman
8. tree top group
9. moth force reduction
10. computer business party
11. hotel model introduction
12. fatigue country house
13. future actor entrance
14. company glass eye
15. concrete desert prayers
16. stone city members
17. margin office development
18. government power semiconductor
19. trout nose money
20. citizen helmet law
21. bath centre bed
22. warrior castle decision
23. software engineering treaty
24. job advance guard
25. tourist growth lakes
26. motorcycle accident burn
27. year marriage lane
28. government hand food
29. dollar bull signals
30. hermit committee report
31. family house structure
32. disaster bread crumb
33. shareholder pressure dispute
34. mother butter knife
35. lightning country debt
36. port city members
37. margin office desk
38. hydrogen bomb explosion
39. party cab driver
40. discharge control law
41. signatory country sugar

42. assistance assembly worker
43. hermit committee decision
44. enterprise product division
45. solvency margin office
46. car exhibition dish
47. steam iron area
48. soya office complex
49. child church bus
50. winter sports creations
51. jungle range control
52. line extension magazine
53. mining donor list
54. soul food firm
55. fishing ground manager
56. post field day
57. day exercise disease
58. discharge control product
59. pole donor list
60. job advance word
61. drug research expression
62. applicant country machine
63. interest man marker
64. exhibition competition law
65. playing surface period
66. summer heat height
67. city government power
68. management control office
69. frog ring worm
70. steel blanket months
71. deficiency control official
72. lightning country hideaway
73. food firm note
74. retirement practice squad
75. part subject deletion
76. domain army personnel
77. herb garden study
78. sightseeing bus function
79. soya office staff
80. box top group
81. rice eye level
82. door control officer
83. heat height limit
84. animal fatigue country
85. country sugar content
86. cigarette helmet law
87. aerospace project manager
88. budget draft food
89. charity case law
90. college sports price
91. student power analysis
92. pole donor fatigue
93. leg exercise disease
94. wax area map
95. search speed trap
96. extension magazine subscription
97. laptop part subject
98. blood bath centre
99. security blanket months
100. candy eye color
101. food assistance base
102. waste application generation
103. aircraft accident dispute
104. bath water control
105. career practice facility
106. drinking water control
107. furniture factory lecture
108. hand food system
109. dawn air lodge
110. state hand charge
111. draft plant project
112. family home jar
113. party hand charge
114. machine core microprocessor
115. student body instrument
116. dualist service enterprise
117. machine core lad
118. hydrogen bomb expert
119. water surface colonies
120. cash assistance base

121. state assistance assembly
122. top group conflict
123. bath water cure
124. gas future officer
125. emergency assistance assembly
126. volume growth lakes
127. music career practice
128. college committee chairmanship
129. refugee child church
130. computer work contract
131. strip search speed
132. water control policy
133. soybean oil ring
134. phantom nation mode
135. canine oil product
136. week extension magazine
137. acting fair official
138. student paper cup
139. development body instrument
140. review identity theft
141. family picture procedure
142. retirement practice test
143. factory lecture series
144. decision development work
145. airport lounge provision
146. government pressure dispute
147. disease imports shot
148. resource revenue collection
149. pet cake shop
150. siege door knocker
151. year paper attempts
152. air pressure dispute
153. student paper trail
154. testtube height restriction
155. nomination paper attempts
156. mother butter texture
157. power analysis group
158. water surface period
159. search speed skating
160. language construction phase
161. gambling machine actor
162. velvet string victim
163. moth force structure
164. sugar factory lecture
165. banking world policy
166. retirement home jar
167. deficiency committee Republican
168. review identity paper
169. divorce paper attempts
170. hand food poisoning
171. post field officer
172. draft food assistance
173. candidate city engineer
174. van factory lecture
175. exercise disease germ
176. filename part decision
177. donor organ area
178. student pressure tactic
179. oil ring worm
180. door control department
181. labour government bars
182. highway assistance base
183. air disaster bread
184. telecommunication policy limb
185. disaster bread pudding
186. family analysis firm
187. labour party cab
188. sports price signal
189. sea machine damage
190. assistance base closing
191. dinner officer corps
192. soft drink room
193. budget picture shape
194. diamond mining donor
195. dualist service worker
196. starvation shock value

197. boating accident dispute
198. ton ship members
199. surgeon fatigue country
200. playoff picture procedure
201. sports body manager
202. construction job advance
203. canine oil deal
204. candy cigarette helmet
205. budget pork wall
206. drink room temperature
207. wage growth lakes
208. birth mother requests
209. storm wind administration
210. aerospace project finance
211. accident burn unit
212. auto part subject
213. donor country machine
214. iodine deficiency control
215. post field trial
216. crime family analysis
217. snack food firm
218. cocoa exhibition competition
219. manufacture computer work
220. domain army man
221. pork wall curvature
222. beehive machine damage
223. birth mother butter
224. peanut folk architecture
225. rice eye opener
226. chocolate burn victim
227. brewery beer fair
228. jungle paper price
229. merchant bank function
230. government power analysis
231. dualist service network
232. government decision development
233. playing surface colonies
234. computer language construction
235. summit country sugar
236. policy limb function
237. party government structure
238. research body instrument
239. farmyard waste application
240. mining donor heart
241. award dinner officer
242. government bank function
243. union family analysis
244. farmer research contract
245. hour siege door
246. cruise ship members
247. poker player industry
248. vehicle part decision
249. cotton order backlog
250. draft plant variety
251. charity report language
252. dawn air work
253. engineering treaty negotiation
254. tablespoon sugar bars
255. jungle paper plate
256. fruit drink room
257. cathedral administration chief
258. down air work
259. chain concession speech
260. nut engineer joke
261. oil interest man
262. engineering student pressure
263. gasoline margin office
264. book review identity
265. company glass product
266. state government bars
267. accident burn victim
268. candy business party
269. sham trust money
270. part decision table
271. fishing pole donor
272. plastic chain concession
273. acting fair play
274. brick home jar

275. world policy wonk
276. extension magazine rack
277. enemy cure rate
278. daisy baby formula
279. health research expression
280. heat height restriction
281. health care party
282. future officer corps
283. troop home jar
284. party hand food
285. army post field
286. charity case discount
287. testtube height advantage
288. cathedral administration building
289. security arrangement fee
290. player industry excutives
291. immigrant child limb
292. core job advance
293. budget draft plant
294. deficiency committee approval
295. lightning country butter
296. jungle force reduction
297. bear helmet law
298. liver disease imports
299. television domain army
300. committee decision development
301. chain concession contract
302. donor government bars
303. motion picture procedure
304. marketing job paper
305. party machine actor
306. college limb function
307. family dinner officer
308. investment account consultant
309. jury fatigue country
310. wage control office
311. cathedral performance measure
312. pole donor conference
313. sham marriage lane
314. hill country sugar
315. counseling student pressure
316. fence post field
317. horse starvation shock
318. horse war criminal
319. pole donor community
320. month extension magazine
321. sports price increase
322. food firm decision
323. handlebar assets sale
324. research community stock
325. steam soldier ant
326. protectionist pressure dispute
327. hotel model car
328. decision development trend
329. poster child church
330. pet cake flour
331. insurance world policy
332. oil future officer
333. report language inventions
334. mining donor organ
335. trust money maker
336. iron area study
337. purse string victim
338. draft plant owner
339. organ area study
340. beard alcohol abuse
341. horse starvation problems
342. peasant family analysis
343. artichoke heart members
344. beard alcohol use
345. operator fatigue country
346. fairground machine core
347. day siege door
348. disaster bread knife
349. chocolate burn unit
350. air work place

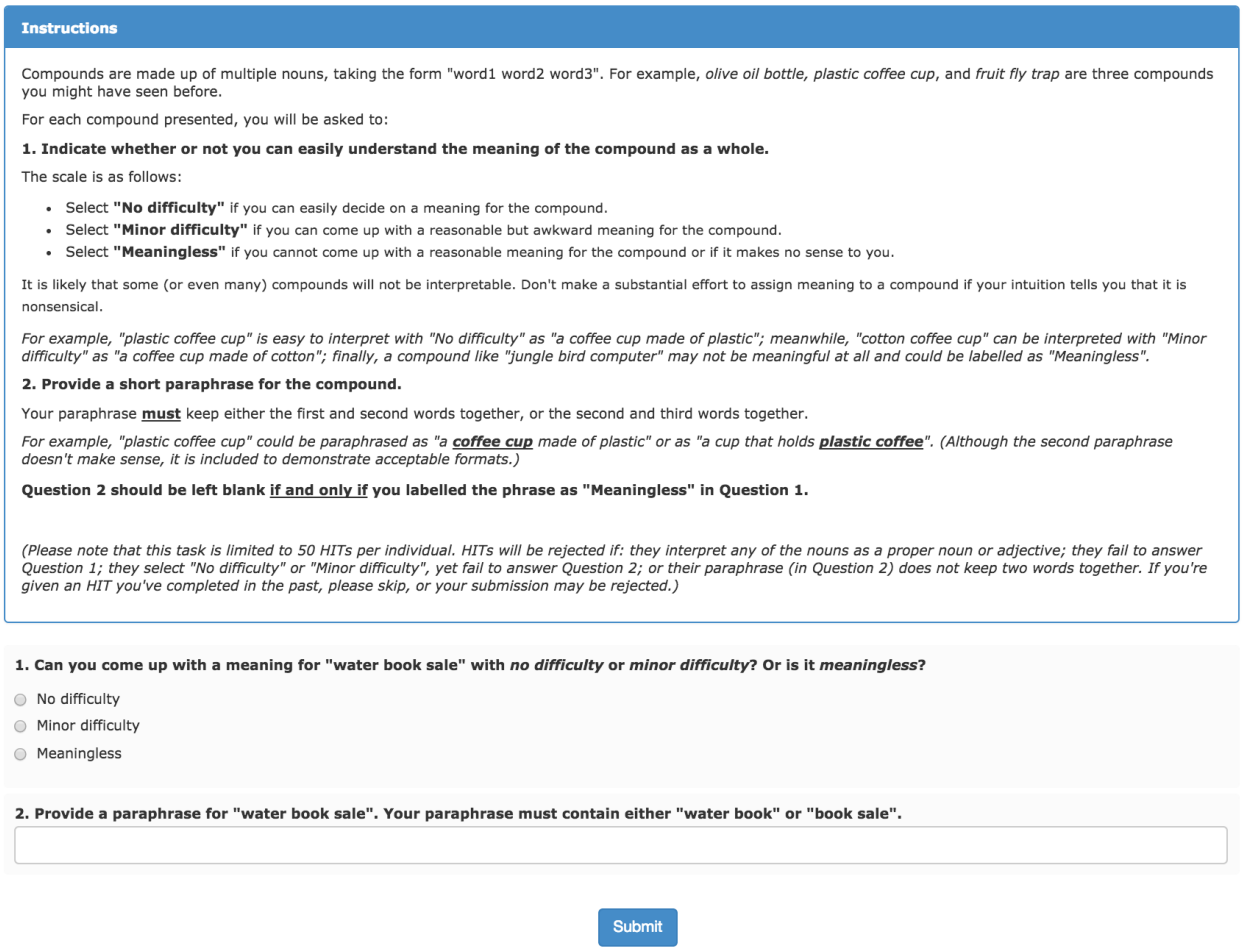
351. enforcement resource revenue
352. assistance assembly plant
353. troop home creations
354. foster child limb
355. daisy baby brother
356. refugee policy limb
357. job advance booking
358. salmon datum centre
359. pitching hand food
360. foot sea machine
361. puppet government power
362. deficiency disease imports
363. peasant party soup
364. training exercise disease
365. land security arrangement
366. minority neighborhood cube
367. child care party
368. sea machine operator
369. killer disease imports
370. country machine dryer
371. draft food supplies
372. motor time value
373. college sports creations
374. pipeline company glass
375. oil future actor
376. assistance base price
377. writing career practice
378. port city dispute
379. draft food business
380. engineering treaty restriction
381. gold mining donor
382. training arena show
383. air work report
384. surplus water book
385. deficiency control factor
386. degree heat height
387. customs post field
388. induction motor time
389. donor heart members
390. foster child fish
391. sap folk doctor
392. peanut folk doctor
393. morning paper attempts
394. water book sale
395. candy eye surgery
396. financing resource revenue
397. world policy matters
398. preadmission review identity
399. part subject area
400. water book industry
401. retirement hotel model
402. charity dinner officer
403. candy eye opener
404. birth disease management
405. body heat height
406. beer fair play
407. jungle force structure
408. soccer body manager
409. apple alcohol industry
410. payroll job advance
411. sensitivity training arena
412. future home jar
413. enterprise product portfolio
414. business college weather
415. drugstore chain concession
416. government power boat
417. business college limb
418. moth force level
419. back part decision
420. minority child fish
421. supermarket chain concession
422. minority student paper
423. line extension friends
424. tablespoons sugar measure
425. daisy baby bottle
426. contract extension friends
427. beard alcohol problem

428. fatigue country sensibility
429. business party activity
430. salmon datum element
431. truck part decision
432. foot mountain utensils
433. country sugar import
434. toilet training arena
435. selling pressure dispute
436. tourist city dispute
437. replacement part subject
438. motion picture shape
439. missile part decision
440. water control product
441. diesel motor time
442. faculty machine translation
443. hog farmer research
444. body instrument noise
445. identity paper attempts
446. player industry publication
447. factory lecture hall
448. air work method
449. state dinner officer
450. capital city engineer
451. care party government
452. blanket months order
453. beehive machine operator
454. minute video example
455. factory lecture circuit
456. teaching church elder
457. copy machine core
458. surplus water cure
459. adventure invasion force
460. input jacket pocket
461. utilization review identity
462. city engineer joke
463. college weather damage
464. resource revenue estimate
465. motor time visitor
466. family analysis team
467. carb air lodge
468. tinfoil blanket months
469. nomination survey report
470. dozen horse structure
471. hair care party
472. system engineering treaty
473. farmer research engineer
474. beard alcohol stain
475. product growth lakes
476. dozen horse war
477. government assistance assembly
478. government body instrument
479. report language construction
480. string victim name
481. child fish pond
482. minority neighborhood lake
483. quality part subject
484. contract extension magazine
485. power analysis branch
486. wind administration office
487. nose money broker
488. exercise disease management
489. nomination survey committee
490. party soup can
491. care party plan
492. family pet cake
493. church teaching friends
494. trust money cost
495. chain concession stand
496. rice eye candy
497. forest resource revenue
498. country machine translation
499. back number plate
500. community stock selling
501. computer work practices
502. country house structure
503. member government statue

504. language construction crane
505. business college committee
506. wax area study
507. soya office construction
508. gardening tool price
509. string victim state
510. child fish farm
511. control office employee
512. sap folk architecture
513. review identity document
514. degree water book
515. immigrant community stock
516. terry arrangement fee
517. freak accident burn
518. township enterprise product
519. marketing job advance
520. deficiency control mechanism
521. fatigue country surrounding
522. extension magazine group
523. student pressure group
524. book exhibition dish
525. fleet headquarters protocol
526. pork wall panel
527. cell growth lakes
528. string victim right
529. top group activity
530. manufacturing job advance
531. lifesaving drug orders
532. food chain concession
533. core mountain utensils
534. wind administration head
535. castle decision table
536. birth disease prevention
537. pig iron area
538. resource revenue bond
539. coalition party cab
540. energy input jacket
541. drink room demand
542. jungle range version
543. waste application fee
544. faculty lake area
545. candy eye shape
546. job paper maker
547. hurricane wind administration
548. cancer drug orders
549. beer fair official
550. child church school
551. ball player industry
552. college weather model
553. refugee child fish
554. business party victory
555. blood sugar measure
556. family pet problems
557. input jacket potato
558. weapon factory lecture
559. garden study participant
560. organ area director
561. ground manager report
562. team bus function
563. faculty machine tool
564. machine actor entrance
565. career practice time
566. siege door frame
567. engineering student paper
568. cottage door control
569. domain army doctor
570. heat height advantage
571. daisy baby anchovy
572. quarterback job paper
573. water cure rate
574. exhibition dish owner
575. iron area map
576. party nomination survey
577. horse war office
578. body manager report
579. grain farmer research
580. starvation shock absorber
581. control office friendships
582. community stock futures

583. watchdog body manager
584. iodine deficiency committee
585. party cab ride
586. export interest man
587. lawmaking body instrument
588. exercise disease prevention
589. cathedral performance appraisal
590. drug chain concession
591. processing factory lecture
592. child limb function
593. college weather scientist
594. research trust money
595. cotton order imbalance
596. automobile part subject
597. college committee report
598. job paper producer
599. kidney disease imports
600. gas future actor

C.2 Sample HIT: Ternary Compounds



Instructions

Compounds are made up of multiple nouns, taking the form "word1 word2 word3". For example, *olive oil bottle*, *plastic coffee cup*, and *fruit fly trap* are three compounds you might have seen before.

For each compound presented, you will be asked to:

1. Indicate whether or not you can easily understand the meaning of the compound as a whole.

The scale is as follows:

- Select **"No difficulty"** if you can easily decide on a meaning for the compound.
- Select **"Minor difficulty"** if you can come up with a reasonable but awkward meaning for the compound.
- Select **"Meaningless"** if you cannot come up with a reasonable meaning for the compound or if it makes no sense to you.

It is likely that some (or even many) compounds will not be interpretable. Don't make a substantial effort to assign meaning to a compound if your intuition tells you that it is nonsensical.

For example, "plastic coffee cup" is easy to interpret with "No difficulty" as "a coffee cup made of plastic"; meanwhile, "cotton coffee cup" can be interpreted with "Minor difficulty" as "a coffee cup made of cotton"; finally, a compound like "jungle bird computer" may not be meaningful at all and could be labelled as "Meaningless".

2. Provide a short paraphrase for the compound.

Your paraphrase **must** keep either the first and second words together, or the second and third words together.

*For example, "plastic coffee cup" could be paraphrased as "a **coffee cup** made of plastic" or as "a cup that holds **plastic coffee**". (Although the second paraphrase doesn't make sense, it is included to demonstrate acceptable formats.)*

Question 2 should be left blank if and only if you labelled the phrase as "Meaningless" in Question 1.

(Please note that this task is limited to 50 HITs per individual. HITs will be rejected if: they interpret any of the nouns as a proper noun or adjective; they fail to answer Question 1; they select "No difficulty" or "Minor difficulty", yet fail to answer Question 2; or their paraphrase (in Question 2) does not keep two words together. If you're given an HIT you've completed in the past, please skip, or your submission may be rejected.)

1. Can you come up with a meaning for "water book sale" with *no difficulty* or *minor difficulty*? Or is it *meaningless*?

- ○ No difficulty
- ○ Minor difficulty
- ○ Meaningless

2. Provide a paraphrase for "water book sale". Your paraphrase must contain either "water book" or "book sale".

Submit