Metaphor paraphrasing and word-sense disambiguation: toward a new approach to automated metaphor processing

Xiaoyu Tong

Abstract

Automated metaphor processing is an indispensable component of natural language processing (NLP) and natural language understanding (NLU). In the past ten years, however, automated metaphor interpretation witnessed less development compared to automated metaphor identification. This study was an interesting attempt to advance automated metaphor interpretation, by looking at a largely overlooked feature of metaphorically used words and building a first-of-its-kind metaphor paraphrase dataset. We addressed the observation of Shutova (2010) that Metaphor Identification Procedure VU University Amsterdam (MIPVU), the most widely acknowledged metaphor annotation procedure, is essentially a word-sense disambiguation (WSD) procedure with an emphasis on metaphoricity. We built a dataset that provides not only apt, literal paraphrases of metaphorical sentences, but also misinterpretations that would result from failed WSD, when a wrong sense is assigned to a metaphorically used word. Through the paraphrases and aptness and literalness annotations included in the dataset, we have found that such misinterpretations are indeed different depending on whether they are based on the basic sense of the metaphorically used word or a non-basic sense. Based on our findings, we propose a suite of NLP tasks that our dataset can be used for, including WSD and a novel way to integrate metaphor identification and interpretation. Our dataset is so far the largest metaphor paraphrase dataset and provides rich metaphor annotation not yet included in any existing metaphor datasets. We believe that our dataset and findings are valuable contributions to the community and will open opportunities to novel approaches to automated metaphor processing.

Keywords: metaphor paraphrase dataset, automated metaphor processing, MIPVU, conceptual metaphors, word-sense disambiguation

1 Introduction

Metaphors are cross-domain mappings in our conceptual systems (Lakoff and Johnson, 1980). For instance, when one says that we shall not be *defeated* by COVID-19, one is probably motivated by an idea that the current situation (the target domain) is comparable to a competition between two people (the source domain), and the disease can be viewed as our rival or enemy. Such conceptual cross-domain mappings are fundamental to our cognitive processes, and their manifestation in language, linguistic metaphors, is pervasive in natural discourse. Consequently, automated metaphor processing has been acknowledged as an indispensable component of NLP and NLU.

Automated metaphor interpretation has mainly been treated as a paraphrasing task (Bizzoni and Lappin, 2018; Bollegala and Shutova, 2013; Mao et al., 2018; Shutova, 2010; Shutova et al., 2012; Su et al., 2017): the mission of a metaphor interpretation system is to 'translate' metaphorical language into literal language, so that the output is directly applicable to downstream NLP tasks such as machine translation. The most common approach that existing metaphor paraphrasing systems take to replace metaphorically used words is to look for words that are used frequently in the same context (Shutova, 2010; Shutova et al., 2012; Bollegala and Shutova, 2013; Mao et al., 2018). This study argues that while this approach can rule out misinterpretations that would arise from failed WSD (i.e., using a wrong word sense to interpret the target word), it does not utilise relations between word senses that are potentially useful for metaphor interpretation.

Consider the manual metaphor identification procedure of MIPVU (Steen et al., 2010). To determine whether a word is used metaphorically, the procedure first identifies the contextual meaning of the word, and then consults dictionaries to see whether a more basic meaning can be identified for the contextual meaning; if so, the word should be marked as potentially related to an underlying conceptual metaphor. In essence, therefore, the senses of a metaphorically used word can be divided into three categories: the contextual sense, the more basic sense, and the non-basic senses. Since the more basic sense and the contextual sense refer to the source domain and the target domain of the underlying metaphor respectively, misinterpreting the metaphorically used word as taking the basic sense is likely to be different from misinterpreting it as taking a non-basic sense.

This study thus aimed to examine the differences between interpretations of a metaphorically used word that presume different senses of the word, and whether the differences would be useful for automated metaphor interpretation. We built a dataset that consists of 864 paraphrases for 201 verb metaphors extracted from the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010), the largest metaphor dataset that employed MIPVU. The paraphrases included apt, literal paraphrases obtained directly through crowdsourcing (using Amazon Mechanical Turk (MTurk)) and semi-automatically generated paraphrases presuming each sense of the target verbs; the latter were based on apt, literal paraphrases for the corresponding word usages, also collected through crowdsourcing. Three types of annotations are presented in the dataset:

1. The metaphoricity of the source texts and the crowdsourced paraphrases was determined using MIPVU. The contextual and basic meanings of the target verbs are also specified, which is not provided by VUA.

2. The literalness of both the source texts and the paraphrases were also annotated by

MTurk workers, who were aware of our definition of literal language but not necessarily had a linguistics background.

3. The paraphrases were also annotated by MTurk workers in terms of whether they were semantically similar to their corresponding source texts.

Our dataset therefore has the following benefits:

1. It is the first metaphor paraphrase dataset that employed the same metaphor annotation scheme consistently for both source texts and paraphrases.

2. It is also the first metaphor paraphrase dataset that reflects how ordinary people (who do not necessarily have a linguistics background) approach paraphrasing tasks. Earlier datasets are primarily based on paraphrases created by linguists.

3. It includes both theory-based and non-expert metaphor annotations, which provides valuable data for the study of human and computational metaphor processing.

4. It is, to the best of our knowledge, the first metaphor dataset that provides contextual and basic sense annotations, which essentially denote the underlying metaphorical mappings instantiated by the target verbs.

Our dataset is still under development. Nonetheless, it is already the largest metaphor paraphrase dataset to date, and arguably involves the most complex source texts compared to earlier datasets, and therefore should be more representative of metaphor use in natural discourse. Most importantly, our dataset employed a systematic approach for producing misinterpretations of metaphorically used words, and, as this report shall demonstrate, suggests a novel method for integrating metaphor identification and interpretation in automated metaphor processing.

In the next section, we review previous research on metaphor annotation and automated metaphor interpretation, as well as the existing metaphor paraphrase datasets. The review leads to the research questions and hypotheses of this study, presented in section 3. We then describe the process of building the dataset (sections 4 to 6). Finally, in section 7, we present and discuss our results and lay out NLP tasks in which our dataset will be useful.

2 Related work

2.1 Metaphor identification

To produce a literal paraphrase of a metaphorical sentence, we first need to determine a method to measure the metaphoricity of any given sentence; otherwise it would be impossible to decide whether the resulting paraphrase only uses literal language. In this study, we adopt MIPVU, which, following Lakoff and Johnson (1980), defines metaphors as conceptual cross-domain mappings and is one of the most widely acknowledged metaphor identification procedure in linguistics.

MIPVU identifies metaphor use on word level. As it regards linguistic metaphors as manifestation of conceptual metaphors, the aim of the method is to identify 'all words in discourse that can be taken to be lexical expressions of underlying cross-domain mappings', which are termed metaphor-related words (MRWs). What are usually called metaphorically used words

are a subtype of MRWs termed *indirect metaphors*. Compare the two metaphorical sentences, (1-a) and (1-b). While the two sentences instantiate the same conceptual metaphor, in which DISEASE is the target domain and ENEMY the source domain, (1-a) is directly related to the conceptual metaphor, with *our enemy* literally referring to a person who is our enemy. It is therefore an instance of *direct metaphor*, the verb *is* serving as a *metaphor flag*, signifying that a metaphor is being constructed. On the other hand, the verb *defeated* in (1-b) refers to an abstract situation of not being able to do something as a consequence of the ongoing pandemic; it does not directly refer to the event of being defeated in a competition. The verb is therefore an indirect metaphor.

(1) a. COVID-19 *is* our enemy.
b. We shall not be *defeated* by COVID-19.

According to MIPVU, a word is marked an indirect metaphor if one can identify a sense of the word that is more basic than its contextual meaning. The more basic sense should be a sense that enters a contemporary dictionary of the language (MIPVU advises annotators to consult the Macmillan dictionary (MD) for English language). It should be separate from the sense that corresponds to the word's contextual meaning, provided that the contextual meaning is also recorded in the entry, and should bear some form of similarity to the contextual meaning. MIPVU further specifies that a sense is more basic if it is more 'concrete, specific, and human-oriented'.

As was pointed out by Shutova (2010), MIPVU can be regarded as a WSD procedure with an emphasis on metaphoricity. Furthermore, that a more basic meaning can be identified for the contextual meaning of a metaphorically used word implies the different relations between word senses with regard to metaphorical mappings. If we see the different usages of a word as nodes in a semantic space and focus on metaphor use, the nodes representing metaphorical usages are only directly connected to their corresponding basic senses; there could be nodes completely separate from each other despite being usages of the same word.

2.2 Metaphor paraphrase datasets

Shutova (2010) created a dataset that contains 62 sets of metaphorical expressions and literal paraphrases. The metaphorical expressions are verb-object and verb-subject constructions selected from a subset of the British National Corpus (BNC) containing 761 sentences. The metaphoricity of the source expressions was evaluated by 3 volunteer annotators who were native English speakers with linguistics background. They followed Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), the predecessor of MIPVU, which had an additional criterion for basic meaning: it should not only be more concrete, specific, and human-oriented, but be historically older as well. It should be noted, however, that the annotators were asked to 'imagine' a more basic sense instead of consulting dictionaries.

The literal paraphrases were provided by 5 volunteer annotators; the instructions were to write down all suitable literal paraphrases for the target verbs. Like the annotators who provided the metaphoricity annotations, these 5 annotators were native English speakers with linguistics background. Nonetheless, the literalness of the provided paraphrases were not examined using MIP. It is therefore questionable whether the source expressions and the paraphrases are compatible in terms of metaphoricity annotation.

Mohammad et al. (2016) released a dataset of 171 pairs of metaphorical sentence and literal

paraphrase. The authors extracted from WordNet example sentences of verbs that had 3 to 9 senses (1639 example sentences for 440 verbs were extracted), and collected metaphoricity annotations of the target verbs through CrowdFlower; each instance was annotated by at least 10 annotators. The authors then created literal paraphrases for 176 of the example sentences; the target verbs in these sentences were considered metaphorical by no less than 70% of the annotators. Literal paraphrases were created by replacing a target verb with a synonym that would make the original sentence literal. The three authors selected synonyms independently and resolved disagreements through discussion; five sentences for which disagreements could not be resolved were discarded, thus resulting in the 171 source-paraphrase pairs.

The dataset created by Mohammad et al. (2016) has the merit of reflecting ordinary people's perception of the metaphoricity of word usages, which is arguably unaffected by metaphor theories. On the other hand, the literal paraphrases reflect the metaphoricity evaluation of the authors; metaphoricity/literalness annotation of the replacements was not crowdsourced. Similar to the dataset created by Shutova (2010), therefore, this dataset can be improved by applying the same annotation scheme to the source texts and the paraphrases.

Bizzoni and Lappin (2018) built a metaphor paraphrase dataset containing 200 sets of 5 sentences; each metaphorical sentence is accompanied by 4 paraphrases with varying levels of aptness. Apart from verbs, the dataset also contains metaphorical usages of adjectives, copula metaphors, and multi-word metaphors.

The metaphorical source sentences, according to the paper, either came from 'published sources' or were manually created by the authors. The authors also manually created the candidate paraphrases. They contended that a variety of possible misinterpretations was considered and stressed the inclusion of opposite interpretations of sentiment related metaphors. For instance, both *I love my job* and *I hate my job* were included as candidate paraphrases of the copula metaphor *My job is a dream.*

Bizzoni and Lappin (2018) collected through MTurk 20 annotations for each pair of source sentence and candidate paraphrase. They employed a unique annotation scheme displayed in fig. 1. It is similar to a 5-point rating scale, from 'strongly disagreeing' to 'strongly agreeing' with whether two sentences can be considered paraphrases.

she cut him down with her words		
she cheered him up with her words	1	Two sentences cannot be considered paraphrases.
she left him bored with her words	2	Two sentences cannot be considered paraphrases, but they show a degree of semantic similarity.
she told him things that made him sad	3	Two sentences could be considered paraphrases, although they present some important difference in style or content (they are not strong paraphrases).
she put him down with her words	4	Two sentences are strong paraphrases.

Figure 1: Example aptness annotation in the dataset of Bizzoni and Lappin (2018)

Compared to the two previous datasets, the dataset created by Bizzoni and Lappin (2018) involve more variations in metaphor use. The inclusion of both apt and inapt paraphrases also

provide richer information for training metaphor interpretation systems. Nevertheless, more information is desirable about the types of misinterpretations included in the dataset and the authors' metaphor annotation scheme or operational definition of metaphor. With regard to the latter information, in particular, the dataset appears to suggest a metaphor annotation scheme different from MIPVU. Consider the example in fig. 1. The strong paraphrase, which is supposed to use literal language, replaces the phrasal verb *cut down* with another phrasal verb, *put down*. While phrasal verbs are treated as single lexical units in MIPVU, the usage of *put down* in the paraphrase seems to be based on a more concrete meaning, of putting a person or object onto a surface, as exemplified in (2). The paraphrase should therefore be marked as metaphorical if we follow MIPVU.

(2) Emma put her bag down and went upstairs. (MD put-down_1 1)

We would like to stress that MIPVU is itself an operationalisation of a particular hypothesis about metaphor and we do not consider it the golden rule for metaphor annotation. However, the existence of competing metaphor theories and annotation schemes also implies that the results of using different metaphor annotation schemes are not necessarily the same, and that the reliability of a metaphor dataset is partly related to the metaphor theory or annotation scheme it adopts. We therefore consider it highly important for building metaphor paraphrase datasets to be specific about the underlying assumptions about metaphor and follow a corresponding metaphor annotation scheme strictly and consistently throughout research.

2.3 Metaphor paraphrasing systems

Shutova (2010) defined automated metaphor interpretation as a paraphrasing task and proposed the first computational system that outputs literal paraphrases of verb metaphors (which were verb-subject and verb-object constructions in which the verb is used metaphorically). The system paraphrases metaphorical expressions by replacing the metaphorically used verb with another verb that should make the expression literal. Given an annotated metaphorical expression, the system first collects candidate interpretations or replacements by searching BNC for words that co-occur with the same context as the target verb. The system then ranks the list of candidate interpretations according to their likelihood of co-occurring with the context and filters out unlikely or figurative interpretations based on WordNet hypernym/hyponym taxonomy and selectional preference distributions. Finally, the first interpretation that remains in the ranked list is used to produce a literal paraphrase of the given metaphorical expression. Shutova (2010) recruited volunteer annotators to evaluate the rank 1 paraphrases before and after the re-ranking based on selectional preferences; a paraphrase should be marked correct if it was literal and had the same meaning as the metaphorical source expression. The system reached an accuracy of 0.81 when evaluated against these correctness annotations.

Building upon Shutova (2010), Shutova et al. (2012) and Bollegala and Shutova (2013) proposed the first fully unsupervised metaphor interpretation systems. The general idea was the same: use the rank 1 interpretation in a ranked and filtered list of candidate interpretations to replace the target verb. Shutova et al. (2012) employed a vector space model to produce candidate interpretations, thus overcoming the limitation of relying on WordNet hypernym/hyponym relations. The system was again evaluated against correctness annotations and was found to reach a precision of 0.52. Bollegala and Shutova (2013) expanded the initial list by searching the Web instead of BNC for candidate replacements, using lexico-syntactic

patterns as queries. The system, evaluated against correctness annotations, achieved a precision of 0.42.

Mao et al. (2018) proposed a system that deals with metaphor identification and interpretation at the same time. Given a target word in a sentence, the system searches WordNet for a synonym or direct hypernym of the word that is more similar to the context words than the target word. If such a word is found, the target word is marked a metaphorically used word, and the found word can be considered an interpretation of the target word. The study also demonstrated that metaphor paraphrasing improves the performance of machine translation systems. Note, however, that the system does not filter out figurative replacements like earlier systems (Bollegala and Shutova, 2013; Shutova, 2010; Shutova et al., 2012). Metaphorical usages can be the most frequent usages of a word, which would lead to a high semantic similarity between a metaphorically used word and the context words. It is therefore questionable whether the output interpretations of this system are indeed literal.

Bizzoni and Lappin (2018) trained a neural model that detects the most apt paraphrase of a given sentence from candidate paraphrases provided by their metaphor paraphrase dataset. Although the system does not generate paraphrases by itself, provided that an apt, literal paraphrase of a metaphorical sentence already exists in natural discourse, the system has the potential of locating suitable paraphrases in a large corpus.

3 Research questions and hypotheses

The output paraphrase of a metaphor paraphrasing system needs to meet two basic requirements: *1)* providing correct interpretation of the given metaphor, and *2)* using literal language. Correctly interpreting a metaphorically used word is essentially a process of WSD: the interpretation should reflect the sense that the target word takes in the given context. The majority of previous studies adopted a distributional semantics approach to this WSD problem, narrowing down candidate interpretations to words that co-occur with the same context (Shutova, 2010; Shutova et al., 2012; Bollegala and Shutova, 2013; Mao et al., 2018).

While we acknowledged the validity and effectiveness of this approach, we were interested in the misinterpretations that would result from failed WSD, the type of misinterpretations that a distributional semantics approach should easily rule out. Just as MIPVU identifies MRWs by interpreting the contextual and basic meaning of each target word, automated metaphor interpretation may also benefit from taking into account the relations between word senses. Given a sentence s that contains a metaphorically used word w, let

- c be the correct interpretation of w, reflecting the contextual meaning of w in s,
- b be the interpretation that presumes the basic sense of w with respect to its contextual meaning, and
- d be an interpretation that presumes a non-basic sense.

Let $s'(c)$, $s'(b)$, and $s'(d)$ be the corresponding paraphrases. Since the contextual sense is by definition related to the basic sense and a non-basic sense in different ways, the inaptness of $s'(b)$ and $s'(d)$ may also be different.

Let us look at the metaphorical usage of *abuse* in the phrase *abuse power*. The three usages of the target verb and the corresponding interpretations are presented in table 1. The

Usage	Interpretation
1: *Prisoners were regularly abused by their guards.*	*b*: *mistreat*
2: *abuse power*	*c*: *misuse*
3: *He was fined for verbally abusing the umpire.*	*d*: *verbally insult*

Table 1: (Mis)interpretations of *abuse* in *abuse power*. The sense numbers and examples in the usage column were extracted from the Macmillan dictionary. The interpretation column provides the type of an interpretation (*b* for basic, *c* for contextual, and *d* for non-basic) and the corresponding replacement for *abuse* in *abuse power*.

$s'(c)$, *misuse power*, is undoubtedly the only apt and literal paraphrase. The $s'(d)$, *verbally insult power*, changes the meaning of the original phrase and could be considered metaphorical or nonsensical. The $s'(b)$, *mistreat power*, is not a suitable paraphrase either. Unlike $s'(d)$, however, the inaptness of $s'(b)$ is mainly due to its metaphoricity; it is more similar to the original phrase in meaning than $s'(d)$.

This study thus aimed to investigate to what extent the difference between $s'(c)$, $s'(b)$, and $s'(d)$ illustrated above is applicable to other words. We hypothesised that while only $s'(c)$ is an apt, literal paraphrase, $s'(b)$ and $s'(d)$ are different in terms of their aptness and literalness:

1. $s'(b)$ is just as apt as $s'(c)$ whereas $s'(d)$ is an inapt paraphrase;
2. $s'(b)$ is equally or more figurative than s, and therefore also more figurative than $s'(c)$, whereas $s'(d)$ tend to be considered nonsensical rather than either literal or figurative.

Testing theses hypotheses required a set of single-word metaphors (i.e., texts in which a single word is used metaphorically) and the above-mentioned three types of paraphrases of these metaphors. Since such paraphrases were not readily available in any of the existing datasets, we built and will make publicly available a new metaphor paraphrase dataset. The dataset contains the three types of paraphrases specified above for 201 single-word metaphors extracted from VUA (rendering 864 pairs of source text and paraphrase in total). We focused on the metaphorical usages of 50 verbs in this study, but the dataset is easily expandable by running the same process for the other verb metaphors we have prepared, as well as the metaphorical usages of words of other parts of speech.

4 Collecting interpretations for each word sense

At the core of our dataset is the metaphor identification procedure of MIPVU and the VUA dataset built by the contributors to MIPVU themselves. We extracted metaphorical source texts from VUA and collected $s'(c)$ paraphrases of the texts through MTurk. The $s'(b)$ and $s'(d)$ paraphrases were created by substituting a metaphorically used word (w) with its interpretations in different contexts (i.e., b and d). The b and d interpretations were also obtained through the paraphrase task we posted on MTurk, using example sentences of each sense of w sampled from MD.

4.1 Preparing source texts

Metaphorical source texts We extracted metaphorical sentences from VUA and edited them so that each source text contains a single metaphorically used word. Apart from focusing on the metaphorical usages of verbs, we set aside the metaphoricity of function words due to their high conventionality and frequency in language use.

To begin with, we extracted from VUA sentences that *1)* contain a verb that is tagged as an indirect metaphor, and *2)* the verb is the only content word with metaphoricity annotation (including both `mrw` and `mflag` tags) in a context window of size 5. This resulted in a total of 2814 instances (sentence-lemma pairs) of the metaphorical usages of 741 unique lemmas that are tagged as verbs in the corpus.

Since interpretations of the lemmas in different contexts were to be obtained using example sentences or phrases sampled from MD, we filtered out lemmas for which MD did not have a verb entry, as well as lemmas having senses that were not exemplified. A total of 1985 instances for 491 unique lemmas remained after this step.

The number of metaphorical sentences per lemma ranged from 1 to 148 ($M = 4.04$, $SD = 11.78$). Since automated metaphor interpretation is usually more interested in novel metaphors, we filtered out lemmas having more than 16 instances, resulting in 1178 instances for 477 lemmas. The remaining sentences were then edited so that each would instantiate the metaphorical usage of a single word. The general principle for editing sentences was to retain the original text as much as possible: for each content word that is tagged as an MRW but is not the verb to be interpreted, the smallest possible sequence of words containing the content word (i.e., the word itself, a phrase, a clause, or more) was removed, so that the remaining text is grammatically acceptable, and the target verb is the only content word tagged as an MRW in the text.

Dictionary examples We selected a maximum of 3 examples for each sense of the lemmas. For senses that were exemplified by more than 3 sentences or phrases, 3 of them were randomly selected; otherwise, all the examples were included.

4.2 The paraphrase task

We asked MTurk workers to paraphrase a source text by changing nothing else in the text except for the given target word (i.e., the verb for which the source text exemplifies a certain usage). We included a script in the Human Intelligence Tasks (HITs) which produced a warning message when it discovered that a provided paraphrase may not meet our requirement about target words. We also required the paraphrase to be apt, literal, and grammatically acceptable. In addition to providing a paraphrase, workers were also asked to rate the aptness and literalness of the paraphrase and the difficulty of the task, on a scale from -2 (strongly disagree with a given statement) to 2 (strongly agree). The aptness and literalness ratings were intended to check whether the workers were aware of our aptness and literalness requirements. The difficulty rating, which was accompanied by a question that prompted workers to briefly explain their ratings, was for finding out source texts that were particularly hard to paraphrase properly or given the requirements. Alternatively, workers could choose to not provide a paraphrase and explain directly why the task was impossible to complete.

The paraphrase task described above gradually took shape through a series of pilot studies,

which we present below.[1]

4.2.1 Pilot studies

We designed a small-scale experiment to answer the following questions before releasing the full list:

1. Should the subsenses in Macmillan entries be treated as distinct from the main senses? The manual for MIPVU indicates that Macmillan subsenses should not be considered as sufficiently distinct from their corresponding main senses (Steen et al., 2010, p. 37). However, Macmillan describes subsenses as 'closely related' to the main senses; they are defined as subsenses 'so that the connection is clear'[2]. We hoped to examine through a pilot study whether interpretations of subsenses are systematically different from their corresponding main senses. If so, we would need to obtain different interpretations for a main sense and its subsenses.

2. Would the complexity of metaphorical relations between word senses influence the aptness and literalness of $s'(b)$ and $s'(d)$ paraphrases? A word that can be used metaphorically may have *1)* a single metaphorical usage, *2)* multiple metaphorical usages sharing the same basic sense, or *3)* multiple metaphorical usages based on different literal senses. We hoped to determine before the full experiment whether such difference should be included in our hypotheses.

3. Which source texts are likely to be difficult to paraphrase and therefore would need further editing? In particular, we were concerned about source texts that *1)* were not a full sentence, or *2)* were unlikely to provide sufficient context for paraphrasing (e.g., using pronouns as the subject or object of the target verb).

Reference items for pilot studies The principle for selecting reference items was to avoid removing from our study any conceptual metaphor instantiated by the sentences extracted from VUA. Target words that only have one instance from VUA were thus not considered. The principle also suggests that novel metaphors should not be removed, as they are less presented in VUA and are particularly valuable for our study. The metaphorical usage of *spend* in the following sentence, for example, has not entered the Macmillan dictionary:

(3) The Department of the Environment alone by 1988 was *spending* half a billion pounds out of an Action for Cities total, involving most Home departments, estimated at £3 billion.

This usage is instantiated by only one candidate source text from VUA. We therefore excluded this reference item from the pilot, so as to avoid removing this novel metaphor from our study.

In addition to the principle, we also excluded sentences from VUA whose edited versions were full sentences that are likely to provide sufficient context for paraphrasing. For VUA instances of the same word sense, therefore, we selected those that may cause difficulty in paraphrasing because of not being a full sentence and/or not providing sufficient context

[1] Examples of different versions of the task can be found in the supplement folder or `https://github.com/xiaoyuisrain/rp2-get-int-html`.

[2] `https://www.macmillandictionary.com/learn/dictionary-entry.html`

for paraphrasing. The counterparts of these instances (reference items that are unlikely to cause difficulty in paraphrasing) were therefore Macmillan examples of the same word sense. As Macmillan examples were used directly as reference items (as opposed to sentences from VUA, which were edited in various ways to make sure that the reference items were sentences or phrases containing only one metaphorically used content word), their paraphrases collected in the pilot could be used together with data collected from the official HITs to create interpretations of target words.

Each (sub)sense would be instantiated by 1 to 3 Macmillan examples. To include more variety in the pilot tasks while keeping the sample size down, we selected reference items from target words that have no more than 5 definitions in the Macmillan dictionary. Since we would also like to avoid excluding novel metaphors from the official tasks, we ignored target words that only have one definition (MIPVU is only concerned with contemporary metaphor use. If a target word only has one contemporary usage in the dictionary, the metaphorical usage must be novel). The number of definitions of the target words to be included in the pilot tasks would therefore range from 2 to 5.

We sorted target words in each number-of-definitions category according to their number of occurrences (as target word) in the candidate sentences extracted from VUA. Starting from the most frequently occurring target word in each number-of-definitions category, we collected data about *1)* whether the word has exemplified subsenses, *2)* the number of its metaphorical senses, *3)* the metaphorical usage instantiated by each associated reference item, and *4)* potential difficulties for paraphrasing each reference item. We examined the same number of target words in each of the number-of-definitions categories until we obtained 3 target words in each of the following categories:

1. The entry of the target word has exemplified subsenses.
2. The entry does not have any exemplified subsense.
3. The target word has a single metaphorical sense.
4. The target word has more than one metaphorical sense.

The 6 target words that fulfilled the above categories only included one reference item that is neither likely to provide sufficient context nor a full sentence. We therefore added another two reference items which instantiated other target words. The selected target words and reference items are summarised in table 2. A total of 58 reference items were selected, including 43 metaphorical items and 15 literal items.

It should be noted that for all the selected target words that have multiple metaphorical senses (*spend*, *link*, and *reflect*), the metaphorical senses are based on the same literal sense of the respective words. A word that involves multiple source domains should have a minimum of 4 senses (2 metaphorical senses and 2 literal senses). Among the 8 such words we have examined, there was only one word, *share*, that featured different metaphorical usages based on different literal usages. Target words of this kind might be sparse in candidate sentences extracted from VUA. We therefore decided to leave the examination of such words for now.

First batch of reference items We also prepared a subset of the 58 items to test whether stricter controls should be introduced, so that a participant would only have access to a single instance of a target verb. Since the aim of the paraphrase task was to obtain interpretations

Lemma	Subsense	# met senses	# MD items		# VUA items		
			Lit	Met	−sentence	−context	Both
Include	False	1	3	3	1	5	0
Regard	False	1	1	3	1	1	0
Describe	True	1	1	4	0	2	0
Spend	False	2	3	4	2	0	0
Link	True	2	3	5	1	2	1
Reflect	True	2	4	4	1	1	0
Suggest	-	-	-	-	-	-	1
Join	-	-	-	-	-	-	1

Table 2: Summary of items used in the pilot studies

that would be specific to different usages of a target verb, it would be undesirable if participants tended to paraphrase the same target verb in the same way, so as to complete the HITs quickly.

More specifically, using the first batch of items, we intended to compare *1)* paraphrases of different instances for the same sense, and *2)* paraphrases of different senses of the same word. Similar to preparing the pilot list, we avoided including in this batch all instances for the same sense; otherwise we might not be able to answer the questions that this pilot study aimed to answer. Among the word usages included in the pilot list, only sense 1 of *link* and sense 2 of *spend* were instantiated by more than 2 source texts. We therefore randomly selected items from the pilot list so that the first batch would include: *1)* 2 instances for sense 1 of *link* and sense 2 of *spend*, respectively, *2)* 1 instances for sense 2 of *link* and sense 1 of *spend*, respectively, and *3)* 1 instance for each of the other 6 verbs. The first batch thus consisted of 12 items.

Pilot 1 The first version of the pilot study was accessible to all MTurk workers, and we believed that the instructions were clearly enough for any English speaker to finish the task correctly. The aptness question was not included in this version either, as we assumed that no one would provide an inapt paraphrase on purpose.

Responses came in almost immediately after the release of the first batch, but most of them clearly ignored our instructions. We collected 3 responses for each reference item, and only 5 of the 36 responses retained the form and order of the non-target words in the source texts. We therefore decided to emphasise our instructions a bit more before releasing more HITs.

Pilot 2 We included a checkbox at the bottom of the instructions and asked participants to tick it after reading the instructions for the first time. On the main page of the HITs, we stated that our paraphrasing task had special requirements and the instructions must be read beforehand. We also used the textbox for entering paraphrase to specify our expectations: for source text *The Conservative manifesto contained a commitment*, for instance, the input textbox displayed *The Conservative manifesto ... a commitment* as a placeholder.

We received 210 responses[3] from 57 workers, 28 of whom ticked the checkbox to indicate they have read the instructions. Despite the increased number of acceptable paraphrases, we

[3]The capacity of each HIT was still 3, but the first-batch items were posted twice by accident, thus the total of 70 HITs.

found that some of the participants misunderstood the purpose of the literalness question. We contacted these participants and asked for explanation of their literalness ratings. It turned out that they understood a literal paraphrase as a paraphrase that not only uses literal language but is also an apt paraphrase of the source text. As the instructions included a definition as well as examples of literal language, the issue indicated that the checkbox was insufficient for us to measure whether the responses we received were based on acknowledgement and adequate understanding of the instructions.

Pilot 3 We changed the checkbox at the end of the instructions into a quiz to test participants' understanding of the instructions. The quiz consisted of two questions, one about the requirement that their paraphrases should not change any of the non-target words, and the other about our definition of literal language (see supplement file `paraphrase-pilot-3/instance.html`). We made it explicit in the HITs that submissions would be rejected directly if the quiz was not completed or completely poorly.

Apart from the quiz, we also added the aptness question to the HITs. While we believed the instructions were clear enough about the difference between aptness and literalness, the aptness question would make it convenient for participants to express any other concerns they had about their paraphrases apart from the use of literal language: if they provided a literal paraphrase but wanted to let us know they were not entirely satisfied with the paraphrase, they could show it in their aptness ratings instead of lowering literalness ratings or having to write their concerns down as comments.

Since about half of the participants ticked the checkbox in the last pilot study and the quiz should be more noticeable than the checkbox, we believed we were ready to start the full experiment. We therefore released the first batch of the full list and expanded the maximum number of responses per HIT to 5. Unfortunately, only 6 out of 56 participants answered the quiz questions; 94% of the responses had to be rejected directly.

Pilot 4 We started to use system qualification to control the accessibility of our HITs, which was strongly recommended by fellow requesters to achieve more efficient data collection. As recommended, we set the minimum number of approved HITs to 500 and the minimum approval rate to 99%.

To further highlight the instructions and the quiz, we added two questions to the main page, asking whether the instructions was read and whether the quiz was completed respectively. We hoped that the questions would serve as a reminder that the quiz must be completed.

The qualification setting proved to be effective: 21 out of 33 participants completed the quiz; only 13% of the responses needed to be rejected directly. However, the participants' answers to the quiz questions were not necessarily satisfactory: 4 of the participants reached a score of 50% or lower. We also learned later on that our HITs must have been avoided by many qualified workers, as our approval rate was merely 11% when we released this pilot study.

We also noticed that the system qualification did not help much with keeping away robots: we continued to receive submissions which copied the source text or our instructions where a paraphrase should be provided, moved the sliders to the same place for all rating questions, and gave 'None' or a random integer as reason for the difficulty rating.

Pilot 5 We took a participant's advice and constructed the quiz as a qualification test that must be taken before working on our HITs. We also determined a set of acceptable answers: submitted tests were graded automatically, and workers whose answer did not meet our standard were still not qualified. To avoid robots who might be able to pass the test by trying out random answers, we allowed every worker a single chance to take the test; a worker who wanted to retake the test had to speak to us directly.

Only two workers submitted answers to our HITs this time. We found this result surprising, as our approval rate when releasing this pilot study was around 30%, which was already higher than last time. We also noticed that many workers took the qualification test and got a passing score, but did not work on any of the HITs. Since MTurk requesters could only contact workers who worked for them, we were not able to ask the qualified workers what kept them from responding to our HITs.

During this pilot study, we updated the description of our HITs several times to explain our low approval rate and encourage workers to contact us if they had any questions about the qualification test or the HITs. We also emphasised that if their answer was acceptable, the HITs were immediately accessible once they submit the qualification test. These did not result in more submissions, however. We therefore decided to restrict the time that the full list would be live to a week. We also planned to create paraphrases by ourselves if our HITs remained unpopular.

4.2.2 The full study

To better manage the progress of the study, we selected 50 verbs to focus on for this study; the full list was thus reduced from over 4,000 to 767 reference items (242 metaphorical source texts from VUA and 525 example sentences from MD). The 50 verbs consisted of 10 verbs from each number-of-senses category, from 2 senses to 6 senses. For each number-of-senses category, we selected the verbs whose usages were instantiated by the largest number of both Macmillan examples and VUA source texts.

The reference items were randomised and split into 28 batches, so that each batch would contain at most a single instance for each target verb. The HITs stayed live for two weeks, during which time we received 2026 responses to 700 unique source texts; the number of responses to each of the source texts ranges from 1 to 5. Due to time limit, we then removed the unfinished HITs and used the obtained data to create pairs of source text and candidate paraphrase. These source-paraphrase pairs and their annotations formed our metaphor paraphrase dataset.

5 Collecting aptness and literalness annotations

5.1 Preparing candidate paraphrases

Two types of candidate paraphrases were prepared:

1. Literal paraphrases obtained directly via the paraphrase HITs.

2. Paraphrases generated automatically by presuming each sense of a target word and then edited by the author.

Selecting crowdsourced literal paraphrases Candidate paraphrases of type 1 were selected from paraphrases of the VUA source texts provided by MTurk workers. We received 615 paraphrases for 222 VUA source texts. We employed MIPVU to determine which paraphrases were literal enough to be selected. Since the target word is the only content word related to metaphor use in each of the source texts, we determined the literalness of a paraphrase by examining the metaphoricity of each of the content words in a substitution: a paraphrase was selected if all the content words it used to substitute the target word were non-metaphorical. We also followed the MIPVU principle of 'while in doubt, leave it in': we left out of the list of literal paraphrases substitutions whose metaphoricity was difficult to decide; the list only contained substitutions that we were certain were not used metaphorically.

This literalness criterion did not guarantee the exclusion of inapt paraphrases, however: there were paraphrases that were literal enough but might have changed the meaning of the original texts. These paraphrases were selected nonetheless; we then used the obtained aptness ratings to exclude literal but inapt paraphrases.

We also corrected any grammar mistakes in the literal paraphrases if correcting the mistake did not require adding content words. A total of 141 crowdsourced paraphrases were eventually selected.

Determining interpretations for each word sense From the 1411 responses to 478 Macmillan source texts, we determined interpretations for each instantiated usage of a target word, which provided the basis for generating paraphrases of type 2 addressed above.

The principle was that a selected interpretation of a word sense should result in an apt and literal paraphrase of each of the instances of that word sense. For a given word sense, therefore, we first excluded inapt paraphrases of the corresponding instances. In particular, we excluded paraphrases that were clearly based on incorrect WSD. Interpretations used in the remaining literal paraphrases were then selected; whether a paraphrase was literal was determined using the same method as for selecting crowdsourced literal paraphrases. If all of the remaining substitutions contained metaphorically used content words, we accepted the ones involving highly conventionalised metaphor use: the contextual meanings of the metaphorically used content words are the first sense in MD, indicating that the metaphorical usages are the most frequent usages of the words.

There were cases in which the selected interpretations of different senses of a word happened to overlap. While we followed MIPVU and presumed sufficient distinction between different senses identified in MD (Steen et al., 2010), we allowed for such overlaps, as they indicate how ordinary English speakers perceive of different word usages, which is also an interesting phenomenon to be investigated.

Generating and editing paraphrases The selected interpretations were then used to generate $s'(c)$, $s'(b)$, and $s'(d)$ paraphrases, by substituting the target verb in a source text with each of the interpretations identified for each sense of the verb. It should be noted that we prepared these paraphrases unaware of whether they were $s'(c)$, $s'(b)$, or $s'(d)$ paraphrases. The classification was possible only after the annotation of the basic and contextual senses of the target verbs, which was done after the rating task to make sure that our editing of the paraphrases would not be influenced by our hypotheses.

5.2 The rating task[4]

5.2.1 The pilot study

Given a source text and a paraphrase, participants were asked to rate the aptness of the paraphrase and the literalness of the source text and the paraphrase respectively. To prevent low-quality answers which submitted without moving the sliders, we set the default ratings to 'neither agree nor disagree' and asked participants to explain their ratings if the default was chosen. Like the paraphrase task, we also emphasised the difference between our definition of apt paraphrase and literal language in the instructions and required participants to follow our definition of literal language when evaluating the texts.

We tested this design shortly before the third pilot study of the paraphrase task. The source texts were MD examples; the candidate paraphrases were generated using randomly sampled interpretations. Participants' understanding of the instructions was tested using a quiz inside the HITs (see supplement file `rating-pilot/instance.html`. Similar to pilot 3 of the paraphrase task, only 6 out of 37 workers completed the quiz, and 90% of the submissions had to be rejected.

5.2.2 The full study

The 864 pairs of source text and paraphrase were shuffled and split into 32 batches, each containing no more than 3 instances for the same target verb and an additional control item. Batches containing crowdsourced paraphrases were not accessible to participants in the paraphrase experiment. For each pair of source text and candidate paraphrase, we created a HIT to collect a maximum of 5 annotations.

A major change compared to the pilot study was that we randomised the order of source text and paraphrase for each HIT, and asked workers to evaluate whether the two given texts had the same meaning, instead of whether the given paraphrase was apt. This change was due to our realisation that all the source texts used metaphorical language. If we found a worker who always strongly disagreed that the source text was literal, we would not know whether the worker was being honest or trying to finish the HITs fast.

Like for the paraphrase HITs, workers needed to pass a screening test to gain access to the rating HITs (see supplement file `rating-full/screening-test.pdf`). The screening test included the same question about literal and figurative language that was used in the paraphrase tasks. We also used the test to filter out workers who stated they may share their answer in the screening test with other workers: it was crucial that the respondents were able to distinguish literal and figurative language use by themselves.

In addition to the screening test, we reviewed responses after the release of each batch to maintain the reliability of obtained annotations at the highest level possible. Workers who gave incorrect ratings to control items or always give positive or negative ratings were immediately excluded from the experiment.

[4]Examples of different versions of the task can be found in the supplement folder or `https://github.com/xiaoyuisrain/rp2-rt-html`.

6 Contextual and basic sense annotation

The contextual and basic senses of the metaphorically used words in the VUA source texts were annotated by the author. It was based on these annotations that the selected or generated paraphrases were categorised into the three categories: $s'(c)$, $s'(b)$, and $s'(d)$.

The annotation took place after the selection of interpretations that we used to generate paraphrases. The contextual meanings were determined by comparing the source texts with the example sentences under each sense, instead of the interpretations selected for each sense. The selected interpretations and the annotations were therefore not necessarily correlated with each other.

Our annotation of the basic senses followed MIPVU: for each target word in the source texts, we searched in its entry in MD for a sense that is separate from the contextual sense and is 'more concrete, specific, and human-oriented' than the latter. In the event that we could not find a more basic sense, we determined that the target word was not an indirect metaphor[5] and marked the source text literal.

7 Results and discussion

We received 4311 responses to 864 source-paraphrase pairs, as summarised in table 3. There were 4 items involving $s'(d)$ paraphrases that remained unfinished for more than 2 days after the planned end date of the experiment; the 9 assignments awaiting response were therefore discarded. We also filtered out responses provided by workers who appeared to use 3 or less rating patterns in more than 10 responses; no response needed to be filtered out.

The 204 $s'(c)$ paraphrases used in the rating task included potentially inapt paraphrases obtained through crowdsourcing: we only examined the metaphoricity of the replacement words when selecting crowdsourced paraphrases. To test our hypotheses, therefore, we identified a subset of the $s'(c)$ paraphrases that could be considered as 'true' $s'(c)$ paraphrases to be used to test our hypotheses (that is, to be compared with $s'(b)$ and $s'(d)$). Apart from the automatically generated $s'(c)$ paraphrases, the subset only included crowdsourced paraphrases that *1)* were not identical to an automatically generated paraphrase, but *2)* were considered as having the same meaning as their corresponding source texts (i.e., similarity rating > 0) by more than half of the annotators. A total of 189 'true' $s'(c)$ paraphrases were eventually identified.

Perceived aptness and literalness of $s'(c)$ The $s'(c)$ paraphrases, which correctly interpret the contextual meaning of the target verbs, were considered as having the same meaning as the corresponding source texts ($M = 1.16$), $t(944) = 29.89$, $p < 0.001$. They also significantly increased the perceived literalness of the metaphorical source texts ($M = 0.50$), $t(944) = 7.4994$, $p < 0.001$. As $s'(c)$ paraphrases are apt and literal paraphrases by definition, the results of the t-tests demonstrated the reliability of the ratings we collected.

[5]It should be noted that such cases are not necessarily incorrect metaphoricity annotations in VUA. As MIPVU emphasises contemporary language use and MD is updated frequently to adapt to language changes, it could happen that the more basic sense identified by the creators of VUA became obsolete and was removed from the dictionary.

Category	# items	# obs.	Sim. s, s'	Lit. s'	Lit. $s' - s$
Lit.-to-lit.	11	55	1.02 (1.42)	1.00 (1.33)	0.09 (1.93)
Contextual					
All	204	1020	1.04 (1.30)	1.16 (1.28)	0.51 (2.03)
Auto	97	485	1.08 (1.26)	1.11 (1.32)	0.53 (2.09)
Crowd	107	535	0.99 (1.32)	1.20 (1.24)	0.50 (1.98)
True	189	945	1.16 (1.20)	1.16 (1.28)	0.50 (2.03)
Basic	178	890	0.71 (1.50)	0.89 (1.47)	0.27 (2.08)
Non-basic	471	2346	-0.02 (1.63)	0.81 (1.46)	0.26 (2.18)
Total	864	4311	0.40 (1.59)	0.91 (1.43)	0.32 (2.12)

Table 3: Mean (standard deviation in parentheses) similarity and literalness ratings of $s'(c)$, $s'(b)$, $s'(d)$, and literal-to-literal paraphrases. Difference in literalness ratings (calculated for each source-paraphrase pair) ranges from -4 to 4; similarity and absolute literalness ratings range from -2 to 2.

Perceived literalness of the source texts The results showed that the metaphorical source texts were perceived as slightly literal ($M = 0.58$, $SD = 1.59$), $t(4180) = 23.88$, $p < 0.001$. While their mean literalness rating (0.58) is not significantly less than the literal source texts ($M = 0.91$, $SD = 1.49$), $t(55.62) = -1.58$, $p = 0.059$, it is significantly less than the crowdsourced paraphrases ($M = 1.20$), $t(778.46) = -10.45$, $p < 0.001$. Recall that like the literal source texts, the crowdsourced paraphrases included in the rating task were annotated as non-metaphorical using MIPVU. The smaller difference between the metaphorical and the literal source texts was probably related to the small sample size of the latter: the comparison was between 201 metaphorical instances and 4 literal instances. The results thus indicated that the MTurk annotators were able to differentiate metaphorical and literal language. Nonetheless, it was not clear from the ratings themselves whether the annotators were simply less sensitive to metaphoricity than trained analysts or used different criteria for metaphoricity annotation despite the instructions.

Perceived aptness of $s'(b)$ and $s'(d)$ Like the $s'(c)$ paraphrases, the $s'(b)$ paraphrases were also perceived as having the same meaning as their corresponding source texts ($M = 0.71$), $t(889) = 14.03$, $p < 0.001$. However, the two mean similarity ratings are significantly different, $t(1695.6) = -7.14$, $p < 0.001$. Our hypothesis is therefore only partly true: while $s'(b)$ paraphrases can be considered as apt paraphrases, they are less apt than $s'(c)$ paraphrases.

Our hypothesis about the aptness of $s'(d)$ is also partly true. The mean similarity rating of the $s'(d)$ paraphrases (-0.02) is not significantly different from 0, $t(2345) = -0.46$, $p = 0.65$. Nonetheless, it is significantly less than the $s'(b)$ paraphrases, $t(1726.5) = -11.92$, $p < 0.001$. It is therefore possible to differentiate $s'(c)$, $s'(b)$, and $s'(d)$ paraphrases in terms of aptness: $s'(c)$ paraphrases are the most apt whereas $s'(d)$ paraphrases are the most inapt.

Perceived literalness of $s'(b)$ and $s'(d)$ The $s'(b)$ paraphrases were perceived as significantly more literal than their corresponding source texts ($M = 0.27$), $t(889) = 3.91$, $p < 0.001$. However, they were perceived as significantly less literal than the $s'(c)$ paraphrases, $t(1764.8) = -4.16$, $p < 0.001$. Our hypothesis about the literalness of $s'(b)$ is, again,

partly true.

On the other hand, the perceived literalness of the $s'(d)$ paraphrases contrasts with our hypothesis entirely. Like the $s'(b)$ paraphrases, the $s'(d)$ paraphrases were perceived as significantly more literal than their corresponding source texts, $M = 0.26$, $t(2345) = 5.82$, $p < 0.001$. Moreover, the $s'(d)$ and the $s'(b)$ paraphrases were perceived as equally literal, $t(1590) = -1.36$, $p = 0.17$.

Let us look at the literalness ratings of the paraphrases of an instance of the target verb *lead*, presented in fig. 2. The contextual meaning of *lead* in this instance corresponds to sense 4. The more basic sense we identified for this metaphorical usage is sense 1; the other senses are non-basic senses. As expected, both $(d, 2)$ and $(d, 5)$ fail to convey the meaning of the source sentence. However, while we expected $s'(d)$ paraphrases to be nonsensical, $(d, 2)$ happens to make sense, although it changes the meaning of the original sentence almost entirely. As the subject of the sentence, *the analyst*, refers to a person, the literalness ratings would correspond to a metaphor annotation of the sentence that followed MIPVU. It is therefore understandable why this $s'(d)$ paraphrase was considered more literal than the source sentence. If paraphrases like $(d, 2)$ occur frequently, that is, if there is high probability that $s'(d)$ paraphrases happen to make sense, it seems to explain why they were found to be evaluated as more literal than the source texts.

On the other hand, $(d, 5)$ is precisely the kind of $s'(d)$ paraphrase that our hypotheses were based upon. The mean literalness rating (-0.8), however, indicated that the paraphrase was considered non-literal instead of nonsensical. There are two possible reasons for this result. Firstly, our rating task might have discouraged the annotators from providing a rating of 0 ('neither agree nor disagree' that the given text uses literal language). We used sliders instead of radio buttons to collect ratings, so that it was in accordance with the idea that different levels of literalness were on a continuum. The problem with using sliders, however, was that they always had a default value, and MTurk workers could submit responses without moving the sliders at all. Our solution was to set the default value to 0, which was the most neutral, and require explanation if the workers were to submit the default value as their answer. While we indeed received ratings of 0 accompanied by explanations, it is possible that many workers preferred non-zero ratings, as entering explanations would consume more time. Secondly, people may tend to make sense of a text no matter how strange it sounds; they may not be prepared to evaluate nonsensical texts. In future studies, therefore, one could prepare annotators by pointing out explicitly that they would receive nonsensical texts. One could also present 'the text is nonsensical' as a separate category, rather than belonging to a continuum of levels of literalness.

The literalness ratings of the $s'(b)$ paraphrase, $(b, 1)$, however, seems less explainable. The paraphrase substitutes *guide* for the target verb, *lead*. The contextual sense of *guide* in the paraphrase is sense 2a, as in (4-a). Sense 1 can be identified as the more basic meaning for this metaphorical usage, as in (4-b). The respective usages of the two verbs in the source sentence and the paraphrase are similar in that both metaphorical usages are based on a more frequently used literal sense. It is therefore not clear why $(b, 1)$ was evaluated as more literal than the source sentence. We can, however, relate the inexplicable ratings of the $s'(b)$ paraphrases to our finding that the mean literalness rating of the metaphorical source texts is significantly above 0. The ratings of the $s'(b)$ paraphrases seems to attribute more weight to the possibility that the difference between MIPVU and crowdsourced metaphor annotation goes beyond different levels of sensitivity to metaphorical or figurative language. In other

words, the way MTurk annotators evaluate the literalness or metaphoricity of a text may be systematically different from linguists who follow MIPVU or conceptual metaphor theory (CMT) in general.

(4) a. We can guide you through the maze of financial planning.
b. He guided them through the forest.

It should be noted, however, that the range of metaphors that should be recognised and interpreted by NLP systems is still an open question: as NLP systems are not necessarily parallel to the conscious mind or any theory of human language processing, neither human judgement nor MIPVU necessarily delimit the right range of metaphorical language for NLP. Nevertheless, it is premature to disregard in automated metaphor processing metaphors that are not recognised by MTurk annotators but satisfy a theoretical account of metaphor (e.g., the ones that can be identified through MIPVU). We contend that it is desirable to include both theory-based and intuitive metaphor annotations in a metaphor dataset, as our dataset does. Both annotations could be useful for automated metaphor processing, and the comparison of the two could lead to valuable findings about human metaphor processing.

s'	Lit.	Diff. lit.
$(c, 0)$ *encouraging*	1.2	0.8
$(b, 1)$ *guiding*	1.0	1.0
$(d, 2)$ *being more successful than the user in a system ...*	1.6	1.4
$(d, 5)$ *living*	-0.8	-0.6

s: *Initially the analyst does all the work, leading the user towards a system he thinks is right for the business.*

(a) Paraphrases and the source sentence

1	in front of *She led us down the hill.*
2	winning *lead the field*
3	in control *She led the team during the project.*
4	cause *I had been led to believe ....*
5	live *lead a happy life*
6	card game *She led with the eight of spades.*

(b) Usages of *lead* in the Macmillan dictionary

Figure 2: Paraphrases of an instance for *lead* and their literalness ratings. Lit.: mean literalness rating of a paraphrase, range $[-2, 2]$. Diff. lit.: mean difference in literalness rating, calculated for each paraphrase-source pair, range $[-4, 4]$; value above 0 means the paraphrase was rated as more literal than the source text. Paraphrase identifier: type of interpretation (c for contextual, b for basic, and d for non-basic) and the corresponding sense number of the target word. For example, $[b, 1]$ means the paraphrase presumes that *lead* takes sense 1 in the source sentence, which is in fact the basic sense. Number 0 means the paraphrase was obtained through crowdsourcing instead of automatically generated using selected interpretations for each sense.

7.1 Suggestions for using the dataset

While the results of the experiments did not fully correspond to our hypotheses, the three types of paraphrases, $s'(c)$, $s'(b)$, and $s'(d)$, proved to be distinguishable in terms of perceived aptness and literalness. More specifically, an $s'(c)$ paraphrase, which correctly interprets the contextual

meaning of the target verb, tends to be considered the most apt and literal compared to the other two types of paraphrases; an $s'(d)$ paraphrase, which uses a non-basic sense to interpret the target verb, tends to be considered the least apt. The dataset thus provide information for finding apt, literal paraphrases for metaphorical sentences. Given a metaphorical source sentence and one or more candidate paraphrases, a metaphor interpretation system trained on the dataset should be able to predict for each candidate paraphrase whether it is an $s'(c)$ paraphrase.

The dataset can also be used for automated metaphor identification. To begin with, the dataset includes literalness ratings of both metaphorical and literal sentences, and there is significant difference between the two. A metaphor identification system should thus to able to predict metaphoricity using the literalness ratings alone. This difference between the metaphorical source texts and the $s'(c)$ paraphrases also suggests a paraphrasing approach to automated metaphor identification: given a sentence, an apt paraphrase of it, and their literalness ratings, the sentence should be marked metaphorical if the increase in literalness caused by the paraphrasing does not reach a threshold (e.g., 0.5, according to this study). This will not only be a novel approach to metaphor identification, but a novel approach to integrating metaphor identification and interpretation as well.

Furthermore, the dataset can be used for WSD for metaphorically used words. Note that the difference between $s'(c)$ and $s'(b)$ implies the possibility of training a model that could not only predict the contextual sense of a metaphorically used word, but the basic sense as well. Our dataset thus opens the first opportunity for automatic annotation of contextual and basic senses, which will be invaluable to the study of conceptual and linguistic metaphors.

8 Conclusion

In this paper, we have presented our metaphor paraphrase dataset and examined misinterpretations of metaphorically used words that would result from failed WSD. We have come to the conclusion that such misinterpretations are potentially useful for selecting apt, literal paraphrases. We look forward to further improving the dataset and its use in WSD and developing integrated systems of metaphor identification and interpretation.

References

Bizzoni, Y. and Lappin, S. (2018). Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Bollegala, D. and Shutova, E. (2013). Metaphor interpretation using paraphrases extracted from the web. *PLOS ONE*, 8(9):1–10.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press.

Mao, R., Lin, C., and Guerin, F. (2018). Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Mohammad, S., Shutova, E., and Turney, P. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Pragglejaz Group (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Shutova, E., Cruys, T., and Korhonen, A. (2012). Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012*, pages 1121–1130, Mumbai, India.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., and Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Su, C., Huang, S., and Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.