Basic Meaning: The Achilles's heel of metaphor identification

Anonymous ACL submission

Abstract

Basic Meaning (BM) is a fundamental concept in metaphor identification, serving as the reference point against which contextual meanings are compared. Despite its central role in the Metaphor Identification Procedure (MIP) and its extension, MIPVU, little attention has been given to systematically defining and identifying BM, which hinders transparency and reproducibility in both manual and computational metaphor annotation. In this work, we focus on BM itself, proposing psycholinguistically and lexically motivated measures to quantify BM in an objective and replicable manner. We introduce new annotation guidelines that build upon previous metaphor annotation methodologies, demonstrating their impact on annotation consistency. Additionally, we present a novel dataset that highlights the heterogeneity in BM interpretation across studies. Our findings contribute to strengthen the foundations of metaphor-related research by improving the clarity, reliability, and reproducibility of BM annotation.

1 Introduction

The most Basic Meaning (BM) of a word is defined as more concrete in opposition to abstract, more precise, as opposed to vague, more physical or related to bodily action, and etymologically older than other meanings (e.g., the word chicken can be used in two different ways: 'a domestic fowl bred for flesh or eggs' or 'a person who lacks confidence'. In this case the most basic meaning would be the first one.)

The definition and identification of BM is one of the key steps in the Metaphor Identification Procedure (MIP) (Steen et al., 2007), a broadly adopted procedure that has inspired many of the architectures used for Computational Metaphor Identification (CMI). MIP (Steen et al., 2007) and its extension, MIPVU (Steen, 2010), propose the most widely used guidelines toward metaphor annotation.The procedure consists of four main phases:

1. Split the text into different lexical units.
2. Identify the basic meaning of every lexical unit.
3. Identify the meaning of the word in context.
4. If there is a contrast between the BM and the contextual meaning, label the lexical unit as metaphoric.

However, even though the notion of contrast between BM against contextual meaning is considered a decisive factor in labelling a word as metaphoric, little attention has been paid to both manual annotation and computational identification of BM. Metaphor identification and, especially, the choice of BM are rather subjective tasks, which need the expertise and careful interpretation of the annotator. That being said, BM annotation often lacks sufficient transparency, thus hindering reproducibility. For example, in MIP's original paper, only the BM of 11 words is transparently described, and in subsequent datasets used for metaphor identification, the BM of only one word is discussed in detail at most.

To the best of our knowledge, the only remarkable effort involving BM annotation is the work by Maudslay and Teufel (2022), where the authors take a subset of VUAM (Vrije Universiteit of Amsterdam Metaphor) Corpus (Steen, 2010) [1] and annotate the basic and non-basic meanings of 94 words. Nonetheless, their focus is on metaphorical polysemy detection rather than on the inherent complexities of detecting BM in isolation and its impact on subsequent metaphor identification.

In this work, given the lack of previous attention to BM itself despite being a core part of metaphor

[1] VUAM is the largest and widely used dataset for metaphor identification which was annotated using MIP.

identification, we take a step back and focus on BM definition and validation—addressing what has long been the Achilles' heel of metaphor identification—to establish a sound starting point for all possible metaphor-related downstream tasks. Thus, setting BM analysis as our main goal led to the following contributions[2]:

1. We propose a set of psycholinguistically and lexically motivated measures of BM, which are transparent, objective, and replicable (Section 3), while also studying their BM-capturing capabilities (Section 6.1).

2. Building on previous metaphor annotation methodologies, we provide new guidelines and metric-based guidance for manual BM annotation, analyzing their benefits for the annotation process (Section 5).

3. Last but not least, as a natural extension of our BM literature analysis, we present a novel dataset comprising 100 examples gathered from over 500 works citing MIP, illustrating the heterogeneity in the interpretation of Basic Meaning (Section 4).

The paper is structured as follows. Section 2 reviews how two fundamental relationships at the core of this paper have been represented in previous works. Specifically, the relationship between basic meaning and CMI, as well as the nexus connecting psycholinguistics and CMI. Section 3 presents our proposed set of objective psycholinguistic and lexical measures to label BM. In Section 4, both Maudslay-Teufel's and our new dataset are presented. Section 5 details our transparent and replicable guidelines for defining BM. In Section 6, the proposed guidelines and metrics are validated with experimental results. Finally, Section 7 presents the qualitative results of this work, reflected in a discussion of challenges faced during the annotation of BM, along with recommendations stemming from the resolution of difficult cases.

2 Related Work

In this section we first summarize works that have used BM for CMI, and then, we review the works that have used psycholinguistics in CMI.

2.1 CMI and Basic Meaning

Most recent computational models designed for Metaphor Identification use the concept of Basic Meaning (BM) in their neural network architectures. For example, Song et al. (2021) and Choi et al. (2021) hypothesize that basic meaning can be encoded in the static embedding of a decontextualized word. To explore this idea, they compare the embedding of a whole sentence with the isolated embedding of the target word being inspected for metaphoric usage. This method has the problem of static embeddings relying on the most frequent collocations of words, thus representing mostly the most frequent meaning, which as stated by MIP is not necessarily the most basic. Su et al. (2021) and Babieno et al. (2022), assume that a lexical item's most common (first) dictionary definition encodes its basic meaning . Thus, they provide the model simultaneously with the target sentence containing the target word and, the first definition in the Oxford Dictionary. However, this approach is also against the MIP guidelines. Finally, Li et al. (2023) offer the cleanest option. They compare the embeddings of the target word in utterances where they were labeled as non-metaphoric (representing literal usage examples of the target word) with the embedding of the target word in the target context (which represent metaphorical usage examples of the word). Indeed, they obtained state-of-the-art results by refining this notion of basic meaning. With that said, this method suffers from two main drawbacks. The first one is that literal examples need to be annotated. Its other liability resides in its insufficient transparency and its failure to take into account MIP's original criteria of *concreteness*, *physicality*, and *precision* for the definition of BM.

2.2 Psycholinguistically and linguistically guided CMI

There is also a broad research line using psycholinguistic and linguistic features to enhance the metaphor identification process.

Psycholinguistic ratings are relevant in metaphor identification because they connect cognition with language. These ratings are objective measures collected by means of interviews, questionnaires, and sometimes neurophysiological techniques such as electroencephalograms (EEG) that help understand how words are processed and perceived by the brain. Some relevant measures are: *sensorimotor ratings* that explore how words such as 'cook'

[2]Code and data available at: https://anonymous.4open.science/r/BM-4891/

are usually associated with taste and smell while green' is highly associated with sight; *concreteness* 'evaluates the degree to which the concept denoted by a word refers to a perceptible entity' (Brysbaert et al., 2014); *imageability* evaluates how easy it is to portray a mental image of something; and *affectiveness* measures how strongly linked a concept is with emotional cues.

In metaphor literature, some authors have used *visual features* (Shutova et al., 2016; Kehat and Pustejovsky, 2021) or *sensorimotor ratings* (Wan et al., 2020) for BM determination. Such techniques align with MIP for defining BM where BM is more physical and easy to imagine or picture in one's mind. Other authors align more with the *concreteness* feature (Maudslay et al., 2020), whereby basic meaning is said to be more concrete in opposition to abstract. Concerning the *precision* feature, no related studies have been found.

To our knowledge, the most complete work that exploits psycholinguistic and linguistic features in metaphor identification is that of Rai et al. (2016), where the authors' identification model relies heavily on both psycholinguistic features (i.e., *concreteness*, *familiarity*, *imageability*, *frequency*, *affectiveness*, and *meaningfulness* from MRC norms (Wilson, 1988)) and syntactic ones (i.e., *lemmatization*, *part of speech tagging*, *named entity type labeling* and *parsed dependencies*). Their work was developed to test metaphor identification directly. We do expand some of its ideas with language models (e.g., the extension of psycholinguistic norms, for which they used WordNet (Miller, 1995) and describe many limitations of using this method) and apply them to enhance the identification of basic meanings prior to metaphor identification. Further, we introduce a measure of *precision*, which can be defined as "exactness in communicating disciplinary meaning "(Grapin et al., 2019).

In contrast to previous work, we do not only use psycholinguistic and lexical features to identify metaphors, but rather to analyze complexity in manual annotation, and increase its transparency and reproducibility. We use psycholinguistic and lexical features to analyze which aspects were taken into account by annotators to label basic meaning and if these are coherent with the proposed MIP guidelines.

3 Characterizing a Basic Meaning

In this section, we describe the metrics that we propose for measuring basic meaning. We aim at capturing two main dimensions: a psycholinguistic dimension (via *concreteness*, *physicality*, *imageabilty*, and *familiarity* measures), and a lexical one (via our *precision* measure, calculated using semantic taxonomic depth and word information content).

3.1 Psycholinguistic Measures

The psycholinguistic measures that can be plausibly associated to MIP, all of which are measured and reported in sensorimotor datasets, are: 1) *concreteness*, described as 'the degree to which the concept denoted by a word refers to a perceptible entity' (Brysbaert et al., 2014); 2) *imageability*, which 'represents the degree of effort involved in generating a mental image of something' (Scott et al., 2019); and 3) *physicality*, which can be understood as the strength of association between concepts and bodily action (it can include how easy they are to grasp or perceive visually). Moreover, although it was not mentioned in MIP, we propose to also include *familiarity* as an extra psycholinguistic feature to capture whether annotators had chosen a definition as the most basic one because it was intuitively more familiar to them.

Psycholinguistic Norms These kinds of measures are usually stored in psycholinguistic norms, which consist of lists of concepts where each of the concepts is given a rating expressing its degree of *concreteness*, *physicality*, *imageability* or *familiarity*.

There are several of these norms available (e.g., Brysbaert et al. (2014), Pexman et al. (2019), Wilson (1988) or Scott et al. (2019)). However, similarly to what was reported in Rai et al. (2016), psycholinguistic norm datasets do not cover large vocabularies and contexts. To broaden their coverage, we advocate for using static word embeddings, in line with very recent and inspiring work by Flor (2024). We extend their work (Flor (2024) focuses only on extending *concreteness*) by: 1) comparing different methods for extending different psycholinguistic norms (*imageability*, *concreteness*, *physicality*, and *familiarity*), 2) exploring whether word2vec (Mikolov et al., 2013) or NumberBatch (Speer et al., 2017) embeddings worked best for augmenting the norms, and 3) studying whether using just one single norm or an aggregation of the ratings from different norms as an expansion seed led to better norm expansions. To this end, the following sources were ultimately used: Brysbaert et al. (2014) for *concreteness*, Scott et al. (2019) for *imageability*, Pexman et al. (2019) for *physicality*, and Wilson (1988) for *familiarity*. The final selection of models to run the extension on each dataset, along with the validation of the extension and out-of-vocabulary words covered, can be found in Appendix B.

We use our extended psycholinguistic norms to compute a measure for *concreteness*, *physicality*, *imageability*, and *familiarity* for every sense of a target word. To do so, first, the definitions of each sense are lemmatized and stop words are removed. Then, each word in the sense is looked up to retrieve its rating. Finally, the mean of the ratings for every word in the definitions for every feature is computed (See Table 1 for an example).

3.2 Lexical Measures

Precision in natural language can be understood as the "exactness in communicating disciplinary meaning" (Grapin et al., 2019). We propose two ways in which this precision can be measured:

- *Precision* according its taxonomic depth: we compute how deep a word is in a lexical taxonomy under the rationale that the more hypernyms it has, the more details the word encodes in a class. This information can be extracted from English WordNet (Miller, 1995), a key lexicographic resource in natural language processing. We compute *precision* as the depth in WordNet's taxonomy of all the words in a sense definition with respect to the word being defined.

- *Precision* according to its Information Content (IC) (Resnik, 1995): IC quantifies the rarity of a word's meaning based on its probability in a corpus. The more specific a concept is, the higher its information content; so, this measure should address the specificity of meaning in communication. Precision_ic is calculated for every word in the sense's definition with respect to the word being defined using the wordnet_ic function from NLTK [3].

We expect words with higher information content and words further down in the taxonomy (or with most hypernyms above) to be more precise. Again, the mean for all the precision measurements of all the words in the definition is computed (See Table 1).

	flesh	of	a	chicken	used	for	food	mean
Physicality	2	-	-	3.6	3.4	-	3.545	3.136
Concreteness	4.59	-	-	4.8	2.64	-	4.8	4.207
Imageability	3.849	-	-	2.875	3.824	-	5.929	4.119
Familiarity	496	-	-	508	598	-	538	535
Precision	NA	6	NA	10	NA	0	4	5
Precision_ic	NA	0.802	NA	3.337	NA	NA	6.109	3.416

Table 1: Sample of psycholinguistic, and lexicographic measures for the first sense definition of the word 'chicken'. '-' represent removed stop words and NA represent OOV words.

4 Basic Meaning Datasets

As mentioned above, although BM is a core element in MIP, it is rarely reported for a complete dataset. The most remarkable exception is the Maudslay and Teufel (M-T) dataset (Maudslay and Teufel, 2022) (See Table 2), which contains 94 examples of words and 555 annotated basic meanings following the MIPVU (Steen, 2010) guidelines (an extension of MIP with some key differences, such as disregarding etymology). It contains not only the most basic meaning but all possible basic meanings of a word. As a drawback, though, it provides only single words, as opposed to the full sentence in which the word appears (its use context).

Word	Label	Definitions
chicken	1	the flesh of a chicken used for food ()
chicken	1	a domestic fowl bred for flesh or eggs; believed to have been developed from the red jungle fowl ()
chicken	0	a person who lacks confidence, is irresolute and wishy-washy ()
chicken	0	a foolhardy competition; a dangerous activity that is continued until one competitor becomes afraid and stops ()
chicken	0	easily frightened ()

Table 2: Sample from Maudslay and Teufel (M-T) dataset. Note how some words have more than one sense marked as BM.

However, in our BM literature analysis, we observed that BM was particularly susceptible to annotator subjectivity. In this vein, as the M-T dataset would only reflect their authors' views on BM, we opted to broaden the BM analysis and assess how other authors interpreted and annotated it. Thus, we created a novel dataset (Example Compilation - EC) compiling examples taken from our analyzed papers.

To build the Example Compilation dataset, we gathered 100 additional examples of basic meaning annotations (See Table 3) compiled through the inspection of over 500 sources citing the MIP

[3] Accesible at https://www.nltk.org/howto/wordnet.html

original paper (Steen et al., 2007). This dataset contains labels for 879 senses belonging to 100 words. The advantages of this compilation are: 1) it better captures the heterogeneity of annotation as it includes examples annotated by many different authors, 2) it provides words in their context of use, and 3) it contains dictionary definitions matched to their WordNet counterparts. Furthermore, while the M-T dataset provides various basic meanings for one word, this dataset is more strict in the selection of basic meaning and most authors only offer one (the most) basic meaning per word.

word	Sentence	Original BM	Wn definitions	Label
Star	In terms of being a well-known star, they need to be psychologically prepared to resist all that pressure	a very large hot ball of gas that appears as a small bright light in the sky at night	a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior	1
			someone who is dazzlingly skilled in any field	0
			any celestial body visible (as a point of light) from the Earth at night	0
			an actor who plays a principal role	0

Table 3: Sample from our Example Compilation (EC) dataset. The 'Label' column captures the best match between dictionary definitions provided in the original papers and a WordNet definition. The match was done manually.

Our dataset serves a twofold purpose: to complement the M-T dataset and to fill a gap in the resources required for advancing the state of the art in CMI.

5 Annotation guidelines

Given the subjectivity of the original MIP and MIPVU guidelines when defining basic meaning[4], additionally to the creation of transparent metrics (Section 3), we also created new guidelines that aim at making the annotation process of BM more replicable by exploiting our proposed metrics and some other recommendations. In this section, we describe the control annotation guidelines (the original guidelines provided to annotate BM) and our proposed extension. Then, in Section 6.2, we evaluate their benefits in the annotation process.

5.1 Original Guidelines (Control)

The original guidelines read as follows:

The annotator is provided with a set of N words and different definitions per word. Among the different definitions the annotator has to decide, without additional guidance, which of them (per word) has a more basic meaning. More than one definition can be annotated as basic meaning. If its a Basic Meaning write "1" in column "BM", else write "0".

Basic Meaning is:

a) *More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.*

b) *Related to bodily action.*

c) *More precise (as opposed to vague).*

Basic meanings are not necessarily the most frequent meanings of the lexical unit.

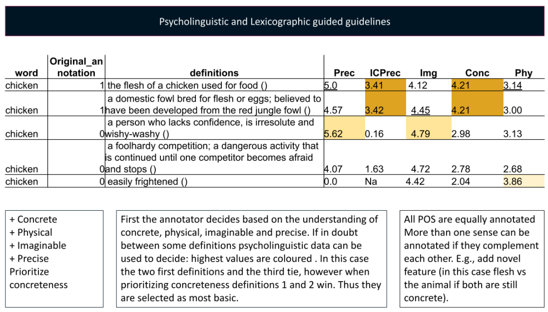
5.2 Guided Annotation

Apart from the original guidelines, our extension provides the following information:

If in doubt between some definitions, psycholinguistic data (mean ratings of precision, imageability, concreteness, and physicality) can be used to decide. If doubt persists, prioritize concreteness.

At this stage, no disambiguation needs to be done, all POS are annotated, and more than one sense can be annotated if it complements another one (adds a new feature).

Moreover, in our annotation process, the annotator will be also provided with the BM metrics values presented as shown in Figure 1, which contains an example.



word	Original_annotation	definitions	Prec	ICPrec	Img	Conc	Phy
chicken	1	the flesh of a chicken used for food ()	5.0	3.41	4.12	4.21	3.14
chicken	1	a domestic fowl bred for flesh or eggs; believed to have been developed from the red jungle fowl ()	4.57	3.42	4.45	4.21	3.00
chicken	0	a person who lacks confidence, is irresolute and wishy-washy ()	5.62	0.16	4.79	2.98	3.13
chicken	0	a foolhardy competition; a dangerous activity that is continued until one competitor becomes afraid and stops ()	4.07	1.63	4.72	2.78	2.68
chicken	0	easily frightened ()	0.0	Na	4.42	2.04	3.86

Figure 1: Guided Annotation: Metrics and visual support for annotators in our proposed extended annotation guidelines.

6 Experimental Results

As validation of the proposed metrics, dataset and annotation guidelines, we conducted two experiments to answer the following research questions:

RQ1 Which features correlate with manual Basic Meaning annotations?

[4]They say little about how annotators should understand precision, relation to bodily experience, and concreteness.

RQ2 Do our proposed annotation guidelines provide any benefit in the annotation process?

To better assess both questions, we split both datasets (EC and M-T) in half, and we made two annotators (A1 and A2) annotate them following the original guidelines (*-Control), and our proposal afterwards in a subsequent step (*-Guided). We followed this setup to focus on the potential benefits that previously trained annotators (i.e., people in the community) would extract from our guidelines. We address the results in the following sections.

6.1 RQ1: Which features correlate with manual Basic Meaning annotations?

The first research question addresses whether prior studies conceptualized BM consistently. Specifically, we examined whether researchers adhered to the notions of precision, concreteness, imageability, and physicality as defined in MIP. To investigate this, we conducted statistical classification experiments using Random Forest models[5] to predict BM based on the metrics outlined in Section 3. The classification models were applied to both the EC and M-T datasets, evaluating their performance in predicting annotations from Various Authors (EC dataset, VVAA), Maudslay and Teufel (M-T dataset, M-T), and Annotators 1 and 2 across both control and guided sets. The model's performance metrics are summarized in Table 4.

	%BM labels	Accuracy	Prec.	Rec.	F1
Random Forest-EC-Guided					
A1	47.4%	67.7	74.1	58.8	65.6
A2	61.3%	81.3	82.5	86.8	84.6
Random Forest-EC-Control					
A1	25.0%	72.9	42.9	16.7	24.0
A2	30.1%	68.1	36.4	21.1	26.7
Random Forest M-T Guided					
A1	49.0%	75.6	68.0	89.5	77.3
A2	47.3%	75.0	85.7	60.0	70.6
Random Forest M-T Control					
A1	25.2%	69.7	0	0	0
A2	31.2%	58.7	16.7	6.7	9.5
Random Forest EC Original annotations					
VVAA	13.6%	87.9	0	0	0
Random Forest M-T Original annotations					
M-T	54.1%	65.5	67.5	61.4	64.3

Table 4: Random Forest results using psycholinguistic and lexicographic features to predict BM. All metrics are reported for definitions labeled as BM.

Three main conclusions emerge from our analysis:

- First, the results highlight the inherent variability and subjectivity in basic meaning annotation, particularly in the case of VVAA (Various Authors)—a collective representation of different researchers citing MIP in the EC dataset. The classification model trained on VVAA annotations was unable to predict basic meaning, suggesting a lack of consistent annotation patterns. This finding underscores the challenge of applying BM annotation in the absence of clear, standardized criteria: when a broad range of interpretations is aggregated, the model fails to detect any systematic patterns.

- Second, the introduction of the newly developed annotation guidelines led to a substantial improvement in model performance, suggesting increased consistency in annotator decisions. The F1-score of the classifier rose significantly from the control to the guided phase: in the EC dataset, the model's performance improved from 24.0 to 65.6 for A1 and from 26.7 to 84.6 for A2. Similarly, in the M-T dataset, F1 increased from 0 to 77.3 for Annotator 1 and from 9.5 to 70.6 for A2. These results strongly indicate that annotators adapted to the revised guidelines, leading to more reliable and systematic BM annotation.

- Third, an examination of feature importance (Figure 2) reveals that concreteness is the most influential predictor across all annotators and datasets, with the exception of VVAA. The diversity in VVAA interpretations likely prevented the model from leveraging any specific feature for prediction. Beyond concreteness, different annotators exhibited varying feature preferences: Maudslay and Teufel (M-T) placed emphasis on concreteness, physicality, and imageability, A2 relied on concreteness and precision, while A1 primarily focused on physicality. Notably, familiarity, a feature absent from the original annotation guidelines, did not contribute to the predictive value of the model.

Focusing on concreteness, we observe that both annotators increased their reliance on this feature when using the new guidelines, aligning their annotation patterns more closely with those observed in the M-T dataset. This suggests that the revised guidelines not only enhanced consistency but also

[5]Random Forest classification was performed with an 80/20 train-test split, 500 trees, seed=0, and three features per split.

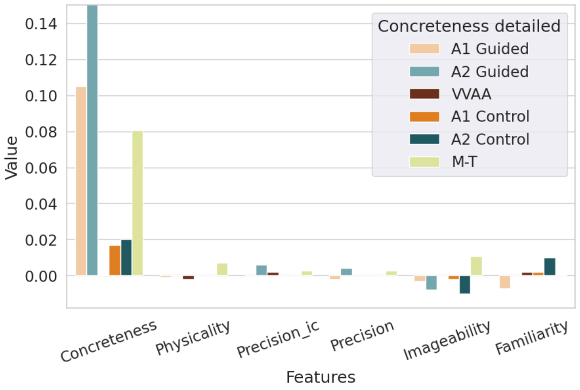


Figure 2: Feature importance in terms of mean decrease accuracy in Random Forest model to predict Basic Meaning. X-axis shows the different inspected features, and each colour represents an annotator. A1 and A2 are split into their ratings in control and guided sets.

encouraged a shared focus on linguistically relevant features.

6.2 RQ2: New guidelines validation

To ensure the guidelines were not only used but useful to obtain better annotations, in this section Inter-Annotator Agreement (IAA) is examined to assess their impact. The IAA results are presented in Table 5, where we report Cohen's Kappa alongside performance metrics (F1, precision, and recall). The inclusion of performance metrics is particularly relevant, as they have been shown to be robust when positive instances (BMs in our case) are significantly outnumbered by negative instances (non-BM) (Hripcsak and Rothschild, 2005).

	Control				Guided			
Annotator	F1	prec	rec	Kappa	F1	prec	rec	Kappa
Maudslay Teufel Dataset (MT)								
A1-A2	64.5	72.0	58.3	0.49	88.1	88.1	88.1	0.84
A1-MT	56.7	83.8	42.8	0.36	78.9	88.9	70.9	0.65
A2-MT	70.9	91.6	57.8	0.54	78.1	88.1	70.2	0.60
Example Compilation Dataset (EC)								
A1-A2	58.7	65.2	53.4	0.43	83.3	91.9	76.1	0.66
A1-VVAA	56.4	46.3	72.1	0.46	31.7	20.6	67.9	0.14
A2-VVAA	58.7	44.8	85.2	0.48	30.4	19.0	75.4	0.09

Table 5: Inter-Annotator Agreement. Metrics reported for examples labelled as Basic Meaning=1.

Overall, IAA increased from the control to the guided set in most cases, supporting the effectiveness of the new guidelines. Notably, the agreement between Annotator 1 and Annotator 2 improved significantly: in the M-T dataset, F1 increased from 64.5 to 88.1 and the Cohen's κ from 0.49 to 0.84, while in the EC dataset, the F1 rose from 58.7 to 83.3 and κ from 0.43 to 0.66. These results validate the newly developed guidelines by demonstrating greater consistency between annotators.

However, the IAA with VVAA (EC dataset) remains notably lower, where agreement between A1/A2 and VVAA did not improve. This further underscores the complexity and subjectivity of BM annotation: when multiple, heterogeneous interpretations of BM are aggregated—as is the case with VVAA—annotator agreement decreases significantly. In contrast, the higher IAA with M-T suggests a more uniform interpretation of BM across annotations, this is consistent with the greater model stability presented on the previous Section 6.1.

7 Discussion

Our analysis of Basic Meaning annotation revealed systematic challenges that highlight both theoretical and practical limitations in existing lexical resources and annotation frameworks. These challenges primarily stem from **ambiguities in lexical databases** and **linguistic complexities in word sense distinctions**, which impact the reliability and consistency of BM identification. Addressing these issues is crucial for refining BM annotation guidelines and improving computational models for lexical semantics. This section first examines the most frequent and problematic cases encountered during annotation, followed by a discussion of methodological refinements that enhance annotation clarity and reproducibility.

7.1 Qualitative analysis of challenging cases of Basic Meaning Annotation

A systematic review of annotator doubts revealed several recurring challenges, stemming from both lexical-semantic properties and linguistic ambiguities.

One set of issues arose from the structure of WordNet (WN). In some cases, multiple definitions were concatenated under the same sense, making it difficult to determine which applied, as in: "attack1: launch an attack or assault on; begin hostilities or start warfare"; in other instances, definitions were overly broad, preventing annotators from assessing concreteness unambiguously : "cultivate1: foster the growth of" were it can be referred to something abstract as in personal growth or to something concrete as in the growth of a plant; additionally, some words had incomplete or underspecified definitions that failed to capture their full range of meanings, as in: "nut1: a small..." (truncated or insufficient). Other difficulties came along with linguistic issues. As in MIP and MIPVU the annotators had to deal with words whose different senses pertain to different parts-of-speech as in "drop2 (noun): a free rapid descent by the force of gravity" and "drop3 (verb): the act of dropping something.", annotators also dealt with transitive and intransitive senses of verbs as in "drown1: die from being submerged in water" and "drown2: kill by submerging in water" or metonymy, as in "chicken1:flesh of the chicken" and "chicken2:domestic fowl". A more detailed description of the challenges and decisions taken for each case can be read in Appendix A.

Many of these issues could be mitigated by treating BM annotation and metaphor identification as separate, sequential tasks. Allowing multiple senses as BM, based on concreteness, precision, and physicality, mitigates issues like metonymy and disambiguation. We propose annotating all senses meeting these criteria, regardless of context or metaphorical use. By structuring BM annotation as an independent step, we ensure that subsequent metaphor analysis builds upon a solid and linguistically sound foundation.

7.2 Recommendations for a transparent and replicable annotation of Basic Meaning

To enhance the clarity and replicability of BM annotation, we propose the following methodological refinements:

Explicitly documenting the selected BM definitions: When publishing metaphor datasets, it is crucial to specify which sense(s) were identified as BM. For example, instead of marking *drown* as metaphorical in *"I'm drowning in work,"* it is useful to explicitly state the BM: *"drown = to be submerged in a liquid."* Providing this information ensures transparency in annotation decisions and facilitates future studies.

Using WordNet over traditional dictionaries: WordNet senses are linked to language-independent identifiers, enabling multilingual approaches to metaphor annotation. Furthermore, these identifiers connect to valuable resources such as Framester, supporting frame-semantic analysis in metaphor research.

Separating BM and metaphor annotation into two distinct tasks: Annotators should first determine BM independently of context and metaphorical usage, ensuring a neutral and consistent classification. Only after BM is established should metaphor annotation take place. This division minimizes conceptual overlap and increases annotation reliability.

Leveraging psycholinguistic ratings to enhance annotation consistency: We found psycholinguistics metrics useful 1) for solving doubts when dealing with ambiguous cases, and 2) as predictors which can expose the annotator's biases. Importantly, these ratings should never override annotators' linguistic intuition, but instead serve as a secondary reference tool. Annotators must remain aware that the primary task is to identify basic meanings independently of metaphorical interpretation, while still adhering to the fundamental criteria of BM (precision, concreteness, and imageability).

By adopting these best practices, BM annotation becomes more transparent, replicable, and linguistically grounded, ultimately improving computational metaphor analysis and annotation consistency across datasets.

8 Conclusion

In this work, we addressed the challenge of defining and annotating Basic Meaning (BM) in a systematic, transparent, and replicable manner. To bridge this gap, we proposed a set of psycholinguistically and lexically motivated measures for identifying BM, demonstrating their effectiveness in capturing key BM properties such as *concreteness*, *precision*, and *physicality*. We further developed and validated new annotation guidelines designed to improve IAA by providing clearer decision criteria. In addition, we introduced a novel dataset derived from over 500 works citing MIP, which illustrates the variability in BM interpretation across studies. This dataset stands as a valuable resource for future research, offering insight into how BM has been understood and applied in different research contexts.

By refining the conceptualization and annotation of BM, we aim to lay a stronger foundation for metaphor identification and related linguistic tasks. Our work contributes to increasing the transparency, reliability, and reproducibility of BM annotation, paving the way for more robust computational and theoretical approaches to metaphor research.

9 Limitations

While our findings provide valuable insights into Basic Meaning (BM) annotation, we acknowledge three limitations. These include constraints related to the annotators, language coverage, and reliance on psycholinguistic metrics, which we outline below.

1. **Annotators:** One limitation of this study is the small number of annotators. Only two individuals participated in the annotation process, one of whom was the creator of the guidelines. To ensure the replicability and generalizability of these guidelines, future work should involve a larger and more diverse group of annotators. We plan to conduct further experiments with additional annotators and extend the evaluation to other languages.

2. **Language**: Our study is currently limited to English annotations. When applying the guidelines to other languages, we anticipate encountering language-specific challenges that may require modifications to the annotation framework. Additionally, psycholinguistic norms for languages other than English tend to be less extensive. However, we hope that the norm augmentation approach used in this paper can help expand such resources for other languages.

3. **Metrics:** Finally, our methodology assumes that existing psycholinguistic ratings are reliable and accurately measured. Future research should further investigate their external validity and applicability to BM annotation. Moreover, we aim to explore the adaptation of psycholinguistic norms to multi-word expressions, as compositional meaning may introduce additional complexities that are not fully captured by current word-level ratings.

Acknowledgments

References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Michael Flor. 2024. Three studies on predicting word concreteness with embedding vectors. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 140–150, Torino, Italia. ELRA and ICCL.

Scott E. Grapin, Lorena Llosa, Alison Haas, Marcelle Goggins, and Okhee Lee. 2019. Precision: Toward a meaning-centered view of language use with english learners in the content areas. *Linguistics and Education*, 50:71–83.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Gitit Kehat and James Pustejovsky. 2021. Neural metaphor detection with visibility embeddings. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 222–228, Online. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226, Online. Association for Computational Linguistics.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Penny M Pexman, Emiko Muraki, David M Sidhu, Paul D Siakaluk, and Melvin J Yap. 2019. Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 english words. *Behavior research methods*, 51:453–466.

Sunny Rai, Shampa Chakraverty, and Devendra K. Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 18–27, San Diego, California. Association for Computational Linguistics.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, page 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51:1258–1270.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451.

Gerard Steen. 2010. *A method for linguistic metaphor identification : from MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Gerard Steen, Lynne Cameron, Alan Cienki, Peter Crisp, Alice Deignan, Raymond W. Gibbs, Joe Grady, Zoltán Kövecses, Graham David Low, and Elena Semino. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287, Online. Association for Computational Linguistics.

Mingyu Wan, Baixi Xing, Qi Su, Pengyuan Liu, and Chu-Ren Huang. 2020. Sensorimotor enhanced neural network for metaphor detection. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 312–317, Hanoi, Vietnam. Association for Computational Linguistics.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

A Recommendations for challenging cases

Analyzing the datasets, we could observe some cases that raised questions in the annotators when labeling the data. Below, there is an enumeration of such cases and how the annotators decided to solve them. Some of them stem from the nature of WordNet entries:

1. Too many definitions in one entry:

- attack1: launch an attack or assault on; begin hostilities or start warfare → many options, but since all possible, mark all as BM.
- attack2: (military) an offensive against an enemy (using weapons)→ less precise than one, but adds detail (i.e., weapons), mark as BM too.
- attack3: take the initiative and go on the offensive → less precise and subsumed by the first two ones, don't add, since it goes against precision.

2. Too broad definitions, not enough to see if they are abstract or concrete:

- cultivate1: foster the growth of→ leave empty cell and annotate in comment 'LIOR' (Look In Other Resource).
- cultivate2: prepare for crops→ in this case, choose this one because it is the most precise among the options.

3. Mixed concrete and abstract:

- take1: remove something concrete, as by lifting, pushing, or taking off, or remove something abstract→ leave cell empty and annotate in comment 'LIOR'.

4. Only one definition, incomplete or unable to account for all senses.

- nut1: →leave cell empty and annotate in comment 'LIOR'.

Other difficulties came along with linguistic issues, most regarding polysemy and metonymy.

1. Different Part of Speech: As in MIP, we decided to cross part of speech boundaries, since senses from different parts of speech can provide relevant information.

- buy1 (noun): an advantageous purchase.
- buy2 (verb): obtain by purchase; acquire by means of financial. transaction → annotate both since one is the process and the other the result.
- drop1 (noun): a shape that is spherical and round.
- drop2 (noun): a free rapid descent by the force of gravity.
- drop3 (verb): the act of dropping something.

2. transitive/intransitive: similarly to the case before, but in opposition to MIPVU all senses were annotated.

- drown1: die from being submerged in water.
- drown2: kill by submerging in water.

3. metonymy:

- chicken1: the flesh of a chicken used for food → clear metonymy (part for the whole), annotate both as BM only if it is coherent with *concreteness*, *precision*, *imageability*, and *physicality*. It would be the metaphor annotator's work to then see the metonymy.
- chicken2: a domestic fowl bred for flesh or eggs.

4. complementary definitions, each definition offers a novel and relevant detail..

- crazy1: someone damaged and possibly dangerous→ mark as BM because it implies physical consequences.
- crazy2: affected with madness and insanity→ mark as BM because it is most precise and linked to a medical condition, which is something very physical.
- crazy3: foolish, totally insane →very similar to two but less precise, do not mark as BM.

5. lexicalization:

- depression1: a mental state characterized by a pessimistic sense of inadequacy → The second definition is more physical and imaginable, however both are precise and concrete. Authors decided both meanings are sufficiently lexicalized in language and refer to two different things, therefore both senses were labelled as BM. In the subsequent annotation for metaphors, it would be the annotator task to see wether the first definition is influenced by the second one.
- depression2: a concavity in a surface produced by pressing.

B Psycholinguistic Norms augmentation

Given the number of words in our datasets (both target words and in the definitions) that were not available on the psycholinguistics databases we expanded them using Flor (2024) approach. Below the reader can find Table 6 summarizing the lengths of the datasets and how many out-of-vocabulary words from the definitions in our data were left.

Norm	Words	OOV BM	Feature
MRC	8228	1306	Familiarity, concreteness and imageability
Pexman	5857	1317	Physicality
Brisbaert	39954	416	Concreteness
Glasgow	5553	1385	Imageability, Familiarity

Table 6: Out of Vocabulary words: Words in Maudslay-Teufel dataset (including target words and in definitions) that were not in psycholinguistic norm datasets.

For the augmentation of psycholinguistics features, we used the Support Vector Machine (SVM) Model from Scikit-learn using NumberBatch and Word2Vec embeddings [6]. To choose the best kind of static embeddings for the augmentation for each feature, as well as for checking if the augmentation was aligned with the manual annotation of the psycholinguistic norms, we also computed Spearman's correlation, a 90/10 data split was used to train/test the model.

Norms	Words	OOV BM	Feature	Spearman w2v	Spearman nb17
MRC	8228	1306	Familiarity	80	78
MRC	8228		Concreteness	85	84
MRC	8228		imageability	78	76
Physicality	5857	1317	Physicality	62	64
Brisbaert	39954	416	Concreteness	79	83
Glasgow	5553	1385	Imageability	79	82
Glasgow	5553		Familiarity	69	69
All norm			Physicality	5	5
All norm			Imageability	30	43
All norm			Familiarity	30	43

Table 7: Spearman values of Psycholinguistic norm augmentation

In Table 7, the best embeddings for each feature and dataset are highlighted in yellow. These are the ones finally used to predict the psycholinguistic features in the Basic Meaning datasets. Darkened, the reader can also see the poor results obtained when

[6] We use a Support Vector Regression with a radial basis function with coefficient $\gamma = 0.003$, ϵ=0.1 and regularization parameter $C = 100$

mixing different psycholinguistic norm datasets and then predicting the Out of vocabulary words. Given such results, we decided that it was best to choose just one dataset per feature, and then augment it with the SVM model.