The Role of Statistical Evidence in Child Acquisition of Multiword Verbs

Aida Nematzadeh

January 2010

1 Introduction

Multiword expressions (MWEs) are combinations of words that convey a single meaning that is not generally composable from their component words. Because of this property, they must, like words, be stored in a speaker's lexicon. However, in contrast to individual words, they occur in phrases whose structure can sometimes vary. For example, *give a loud shout* is an acceptable variation of the MWE *give a shout*. On the other hand, unlike literal (compositional) phrases, the syntactic structure of MWEs is often somewhat restricted. The MWE *give a shout* is less acceptable in a topicalization such as *that shout, I gave* (*cf.* the compositional expression *that movie, I loved*). In addition to the meaning of an MWE, speakers must also know its acceptable variations. In other words, multiword expressions have properties of both lexical units and syntactic structures that must be learned by children when acquiring their language.

Because MWEs are intermediate constructions for which both lexical and syntactic properties must be stored in a lexicon, they do not fall neatly into the traditional grammar model of learning syntactic rules separately from learning the words of the language (which are inserted into those rules). Consequently, multiword expressions can play an important role in supporting the usage-based, construction-grammar account of language acquisition which argues that children learn

various types of constructions (from words to various types of phrasal constructions) from exposure to them in the input (Goldberg, 1995; Tomasello, 2003). Specifically, showing that MWEs can be learned from statistics over usages in child data can support this usage-based constructional view.

However, research on computational modeling of child language acquisition has generally fallen into two major groups. The first group focuses on various aspects of word learning (*e.g.*, Frank et al., 2007; Fazly et al.; Tenenbaum and Xu, 2000), while the second group concentrates on grammar learning (*e.g.*, Sakas and Fodor, 2001; Clark, 2001). The work on intermediate constructions is mostly limited to identifying general properties of verb argument usages (*e.g.*, Alishahi and Stevenson, 2008). Consequently, the problem of how children can recognize and learn the meaning of MWEs has not been addressed.

In computational linguistics, on the other hand, there is much research on MWEs, which has focused on statistical methods for identifying particular constructions (*e.g.*, idioms, light verb constructions, and collocations) from a large corpus (*e.g.*, Fazly et al., 2007; Dras and Johnson, 1996; Lin, 1999). However, MWEs have not been studied in the context of corpora of child-directed speech, nor has consideration been given to the types of statistical measures that could be appropriate in child language acquisition. In the work presented here, we investigate possible usage-based cues that children might use to distinguish between various types of MWEs in child-directed speech, focusing on simple cues that might be cognitively plausible for a child to use.

We focus on a group of multiword expressions consisting of a verb and a noun as a direct object of the verb, and refer to them as verb-noun pairs or verb-noun compounds. Verb-noun compounds are frequent in many languages such as Farsi, Italian and English. Generally, such expressions are formed from basic verbs, which are a group of highly frequent verbs expressing physical actions, such as *see*, *give*, *take*, *make*, *put*, and *get* (Cowie et al., 1975). In addition to their core physical meanings, these verbs have metaphorical meanings when they combine with other arguments in a multiword expression (*e.g.*, *give* in *give a shout*).

Verb-noun pairs using a basic verb form a continuum of expressions ranging from completely literal (compositional) to completely non-literal, forming classes of MWEs that have different lin-

guistic properties. Although there is no clear boundary to separate the expressions in this continuum, four important classes can be distinguished by the way the noun and verb contribute to the meaning of the expression (Fazly and Stevenson, 2007). Below, we describe these classes by explaining the semantic contribution of the verb and noun along with an example from child-directed speech data taken from the CHILDES database (MacWhinney, 2000):

1. **Literal combination (LIT):** *give (me) the lion*
    - give: physical transfer of possession
    - NP: typically a physical entity
2. **Abstract combination (ABS):** *give (her) time*
    - give: abstract transfer or allocation
    - NP: often abstract meaning
3. **Light verb construction (LVC):** *give (the doll) a bath*
    - give: convey/conduct an action
    - NP: predicative meaning (*i.e.*, the NP and verb together determine the predicate of the clause)
4. **Idiomatic combination (IDM):** *give (me) the slip* ["evade"]
    - give, NP: highly abstract contribution (or none)

As can be seen, there is a general pattern for the semantic contribution of the noun and verb in each of these classes that may be a cue for children to learn the meaning of a new expression. For example, when a child recognizes *give me the red doll* as a literal expression, she knows to give an object to the speaker even if she cannot understand the meaning of *red* or *doll*. On the other hand, if she identifies *give a shout* as a light verb construction, she might guess this expression means *shout* (the noun has a predicative meaning in this class). In this paper, we determine simple statistical

measures that reflect the various semantic properties noted above, and that could plausibly be used by a child to make these distinctions among the expressions from the different classes.

In Section 2, we give a brief background of related work done on identification of multiword expressions. Section 3 introduces our measures for identifying different types of expressions (as in (1) to (4)), along with their linguistic intuition. In Section 4, we experimentally evaluate our measures and analyze them. The first set of experiments evaluates the performance of each measure individually and the second set of experiments examines the goodness of the measures when they are used together. Finally, in Section 5 we summarize our conclusions and discuss possible directions for future work.

2 Related Work

Existing research on MWEs has focused on the identification of particular combinations (*e.g.*, idioms) from corpora. Additionally, some work concentrates on separating non-literal expressions from literal ones using the available statistical evidence over usages of expressions. This section is organized as follows: First, we go over the current work on extraction of light verb constructions from given corpora. Then, we explain methods for automatic identification of idiomatic combinations. Finally, we look at classification of a list of expressions into different classes such as LVC and IDM.

2.1 Automatic Identification of Light Verbs

Nominalization, a very frequent phenomenon in many languages, is the process in which a verb is replaced by a construction that includes the nominal form of the verb (*e.g.*, *make a proposal* instead of *propose*). The original verb is often substituted with a light verb (support verb), *i.e.*, a verb that has little semantic contribution to the meaning of a construction. The choice of the light verb varies for each verb (noun in the new construction) and is not predictable. Consequently, it presents a challenge to both language learners and natural language processing systems.

Grefenstette and Teufel (1995) introduce a method for automatic identification of light verbs from corpora. Their method takes pairs of verbs and their nominalized forms as input and finds the best possible light verbs for each pair. The final list of possible light verbs for a nominalization can be used by lexicographers.

To find the list of possible light verbs, the authors extract all the sentences containing a morphological variation of verbs or their nominalized forms. Then, each sentence is tagged with part-of-speech labels. One of the problems they face is that many nouns that are used with light verbs are also used as concrete nouns. As a result, the verbs whose direct object are such concrete nouns must be removed from the possible light verb list. For example, *proposal* is used in concrete form in *He put the proposal in the drawer* but as a nominalized verb in *He made a proposal*, so *put* cannot be a correct light verb for it.

However, nominalized forms are expected to keep some properties of the original verbs and have a similar syntactic structure to their original verbs. Based on this, the authors extract all the prepositional phrases following the original verbs and keep the nominalized forms that occur with similar prepositional phrases. Verbs whose direct objects are the remaining nominalized forms make the final possible light verb list. For example, the prepositional phrases that proceed the verb *appeal* often start with the prepositions *to*, *for* and *in*. Thus, the usages of *appeal* as the head of a noun phrase following these prepositions are considered true nominalizations.

In conclusion, Grefenstette and Teufel (1995) use frequency of co-occurrence of a light verb and a nominalization form to find the correct light verb for the nominalization. They also use the similarity between the syntactic structure of the original verb and its new construction (containing the nominalized verb) to remove the incorrect light verbs from the possible verb list. They test their method on ten nominalizations and it finds the correct support verb for seven of them.

Dras and Johnson (1996) present another method for automatic identification of light verbs. They define f_{ij} to be the frequency in which a verb i occurs as a light verb supporting a given nominalization j. The best light verb for a nominalization is then the one that has the maximum f_{ij}.

In order to compute f_{ij} the authors use two terms: m_{ij} and p_{ij}. The term m_{ij} refers to the number of times the verb i occurs with the nominalization j (not necessarily in a light verb construction). They also define p_{ij} as the probability of the verb i acting as a light verb given the nominalization j. The authors calculate f_{ij} as the product of m_{ij} and p_{ij}:

$$f_{ij} = m_{ij} \times p_{ij} \tag{1}$$

However, the term p_{ij} cannot be estimated directly. As mentioned earlier in this section, Grefenstette and Teufel (1995) assume this probability to be equal to one, thus f_{ij} is equal to m_{ij} in their method. Grefenstette and Teufel (1995) also apply a filter to the list generated based on this assumption. Dras and Johnson (1996) use the unconditional probability over all nominalizations (p_i) as an approximation of p_{ij}. This idea is inspired by the demographic model of Pollard et al. (1981) in which the global population rates are used to calculate the statistics on a subpopulation. The p_i is estimated by counting the occurrences of the verb i with all nominalizations (assuming it acts as a light verb in all the cases).

Dras and Johnson (1996) evaluate their method on a set of 15 nominalizations and achieve an accuracy of 93%, which is higher than the accuracy of the method proposed by Grefenstette and Teufel (1995). Consequently, simple frequency-based measures can be used to automatically identify light verbs. However, in current work we focus on identifying different classes of expressions; thus, we cannot directly use these measures, which only find the best light verbs for a given nominalization.

2.2 Automatic Identification of Idioms

Idiomatic combinations are commonplace in various languages. They form a heterogeneous class of expressions with different syntactic structures (*e.g.*, *bright and breezy*, *see a man about a dog*, and *make a killing*). However, all idioms have a degree of non-compositionality, *i.e.*, their meanings can not be understood by combining the literal meanings of their individual words. In addition,

idioms are known to share certain linguistic properties such as lexical and syntactic fixedness. Fazly et al. (2009) introduce an unsupervised method for type identification of idioms based on their linguistic properties.

The authors focus on a group of idioms, verb-noun idiomatic combinations (VNICs), which consist of a frequent verb and a noun as the direct object of the verb (*e.g.*, *keep one's word*). They propose a method for recognizing VNIC types in a mixed set of idiomatic and literal expressions that draws on the linguistic behavior of idioms. In particular, they measure the degree of lexical and syntactic fixedness of each expression to determine its degree of idiomaticity. Degree of lexical fixedness of a VNIC is is related to whether that particular verb+noun combination is highly entrenched compared to expressions resulting from substituting its constituents by semantically similar words (*e.g.*, *build a killing* is not a valid alternative to *make a killing*). Syntactic fixedness quantifies the variety of syntactic patterns that a VNIC takes.

To measure the degree of lexical fixedness ($\text{Fixedness}_{\text{lex}}$), the authors first generate variants of each VNIC by replacing its constituents with similar words (taken from a thesaurus). Then, they examine the idiomaticity of these variants by calculating the strength of association between them using the pointwise mutual information measure or PMI (Church et al., 1991). They define the degree of lexical fixedness of a verb-noun pair as the difference of its PMI with the average PMI of its variants.

In order to estimate syntactic fixedness ($\text{Fixedness}_{\text{syn}}$), the authors first identify the syntactic patterns that VNICs often take. These patterns are composed of a verb (active or passive), a determiner (*e.g.*, *a/an*, *the*, NULL, etc), and a noun (singular or plural). They measure the degree of syntactic fixedness of a target verb-noun pair by comparing its syntactic behavior to a typical verb-noun pair (with respect to the patterns). To do so, a prior probability distribution is calculated for each pattern *pt* ($\text{P}(pt)$), which is the probability that a typical verb-noun pair takes that pattern. Then for each target verb-noun pair, they calculate the posterior probability distribution over all patterns given the target verb-noun pair ($\text{P}(pt|v,n)$). They compare the syntactic behavior of the target verb-noun pair (the posterior distribution) to the syntactic behavior of a typical verb-

noun pair (the prior distribution) using the Kullback Leibler (KL) divergence (Cover and Thomas, 1991). Additionally, the authors combine the lexical and syntactic fixedness (by weighting and adding them up) to make a unified measure ($\text{Fixedness}_{\text{overall}}$).

The authors compare the classification and retrieval performance of their measures with two collocation extraction measures (PMI and the measure of Smadja (1993)) in two separate tasks. Their results show that $\text{Fixedness}_{\text{overall}}$, $\text{Fixedness}_{\text{syn}}$ and $\text{Fixedness}_{\text{lex}}$ perform better than the other measures, and $\text{Fixedness}_{\text{overall}}$ is the best measure among them. They conclude that using $\text{Fixedness}_{\text{syn}}$ and $\text{Fixedness}_{\text{lex}}$ in a single measure improves the performance, because the errors introduced by the individual measures are removed.

Although the measures proposed in this work are suitable in identifying VNICs in a large corpus, they are not apt in the context of child language acquisition. First, because their formulation is complex, it is not reasonable to assume that a child can compute them. Second, they embed rich linguistic knowledge (using a thesaurus in $\text{Fixedness}_{\text{lex}}$ and complex syntactic patterns in $\text{Fixedness}_{\text{syn}}$) that are not available to a child.

2.3 Automatic Classification of Verb-Noun Combinations

Above, we described work on identification of individual kinds of MWEs. Some work on MWEs focuses on classification of MWEs into different kinds of expressions. As an example, we explain the automatic classification of verb-noun combinations by Fazly (2007). The author introduces four classes of verb-noun combinations (LIT, ABS, LVC, and IDM; see Section 1). She explains some of the significant properties of the non-literal classes (ABS, LVC, and IDM) and devises measures that capture these properties.

Fazly notes institutionalization, lexicosyntactic fixedness, and non-compositionality as the properties of the non-literal class. Institutionalization happens when a group of words are accepted as a semantic unit (*e.g.*, *strong tea*) and is a necessary (but not sufficient) property of MWEs. Lexicosyntactic fixedness is the degree of lexical and syntactic fixedness in an MWE and varies for different MWEs. This is neither a necessary nor a sufficient property for MWEs. Non-compositionality, as

we described Section 2.2, occurs when the meaning of a construction cannot be identified from the meaning of its constituents. This property is also neither necessary nor sufficient for a combination to be a MWE. The author uses some measures that draw on these properties as features in a classification task.

She uses two measures to estimate institutionalization: the frequency of the verb-noun combination and the pointwise mutual information (PMI). The frequency of the verb-noun pair is often used as measure of institutionalization, because the verb and the noun of a verb-noun pair behaving as a semantic unit are expected to often co-occur. The PMI depicts the strength of association between the verb and the noun in a verb-noun pair by comparing the joint probability distribution of the verb and the noun to their independent distribution.

To measure the degree of lexicosyntactic fixedness, the author uses $\text{Fixedness}_{\text{lex}}$, $\text{Fixednes}_{\text{syn}}$, and $\text{Fixedness}_{\text{overall}}$ (the same measures as in Fazly et al. (2009)) along with some other measures more specific to LVCs. One of the salient properties of LVCs is their flexibility in verbal modification. This is perhaps the main reason behind their preference over simple verbs. It seems to be easier for speakers to modify LVCs with adjectives than to modify simple verbs with adverbs (*e.g.*, *have a quick/long look* instead of *look quickly/for a long time*). Thus, LVCs are expected to often appear with an adjectival modifier (Akimoto and Brinton, 1999). The author develops the $\text{Fixedness}_{\text{mod}}$ measure, which is formulated similarly to $\text{Fixedness}_{\text{syn}}$, to estimate the degree of adjectival modification. In this measure, the prior probability distribution of modification over all verb-noun pairs ($\text{P}(mod)$) is compared with the posterior probability distribution of modification given a verb-noun pair ($\text{P}(mod|v,n)$).

The degree of non-compositionality of an expression is related to the similarity of the meaning of its components to its own meaning. Highly non-compositional verb-noun combinations are expected to have a very different meaning from the literal meanings of their constituents. To computationally approximate this intuition, the author compares the context of each target verb-noun pair with the context of its components. The context of a word is defined as a vector consisting of words within a fixed distance with their co-occurrence frequency. Similarity of the context of

a target verb-noun pair with the context of each of its components (the verb and the noun) is measured using the cosine similarity measure, *i.e.*, the cosine of the angle between the two vectors (the verb-noun pair and the verb or the noun).

The author uses the above measures along with some other measures (*e.g.*, semantic category of the noun and the verb) as features of classification. A supervised classification (C5.0[1]) is performed on the data taken from the British National Corpus and annotated by native English speakers. The accuracy of classification is 58.3% on a task with a 25% baseline, giving a relative error reduction of 44.4%.

Consequently, using linguistically driven measures can result in a reasonable performance in the classification task. However, these measures are designed to work on large corpora and are not necessarily suitable in the context of child language acquisition. Some of these measures are very simple statistical measures that are cognitively plausible for children (*e.g.*, PMI), while the others are more complex (*e.g.*, $\text{Fixedness}_{\text{overall}}$).

We note that some measures in the current work are taken from or inspired by the work of Fazly (2007). We adjust these measures to fit in the context of child language acquisition, and develop some new measures, all of which are explained in the next section.

3 The Measures

We hypothesize that children are sensitive to the linguistic properties of each class of verb-noun compound. To examine our hypothesis, we devise some measures based on the linguistic properties and evaluate their performance in separating expressions from the different classes. We note that some of these linguistic properties are motivated specifically by known properties of LVC and/or IDM expressions. The ABS class is not as well understood with respect to their linguistic properties, but we note that they often exhibit similar behavior to LVCs. Because there is some overlap in the properties exhibited by the various non-literal classes, we focus on the overall goal of distinguish-

[1] `http://www.rulequest.com/`

ing non-literal expressions (ABS, LVC, IDM) from literal ones (LIT). As we see later (Section 4), there is only one instance of an IDM in our data, so we remove it. Hence, in our presentation of the measures here, we discuss the properties with respect to the ABS +LVC classes as the non-literal expressions.

As discussed in Section 2, existing research on MWEs has focused on developing statistical measures for extracting multiword expressions from large corpora (*e.g.*, Evert et al., 2004; Fazly, 2007; Fazly et al., 2009). In this work, we need to pick or devise measures that are simple enough to be cognitively plausible for children. As we mentioned above, some of the measures explained in this section are taken and adapted for this purpose from Fazly (2007). The resulting measures fit into three groups based on the linguistic properties of the verb and the noun: Association of a verb-noun pair, semantic properties of the noun (in a verb-noun pair), and the syntactic structure of a verb-noun pair.

3.1 Association of a verb-noun pair

In literal expressions, in contrast to non-literal ones, various nouns can replace the noun component of a verb-noun compound. On the other hand, the number of nouns that a verb takes in non-literal expressions is very restricted. The underlying reason for this might be that the noun in a literal expression is a concrete noun that can be replaced with various nouns, while the noun in a non-literal expression is often an abstract noun which cannot be substituted with many other nouns. As a result, non-literal verb-noun compounds are more entrenched and the noun and the verb are expected to co-occur more often compared to literal ones (Evert, 2008). We looked at three different measures related to the frequency of a verb-noun pair which are described below.

3.1.1 Frequency

The simplest way to measure the association of a verb and a noun is by looking at the frequency of co-occurrence of the verb-noun pair, which is:

$$\text{Cooc}(v,n) \doteq \text{freq}(v,n|\,\text{gr} = \text{dobj}) \tag{2}$$

where gr = dobj is the grammatical relationship between the noun and the verb, *i.e.*, the noun is the direct object of the verb. Here, we assume that children are able to keep track of simple counts of such verb-noun pairs.

3.1.2 Conditional Probability

Although non-literal expressions are expected to co-occur more often compared to literal expressions, the co-occurrence of some literal expressions is also significant (*e.g.*, *take the toy* in child-directed speech). However, the noun in a non-literal expression does not occur with as diverse a set of verbs as a noun in a literal expression. For example, *apple* can be used in many literal expressions with different verbs: *give the apple*, *take the apple*, *eat the apple*, and *wash the apple*, whereas *decision* only occurs in one non-literal verb-noun compound: *make a decision*. In other words, while the verb in a LIT expression is typically thought of as **selecting for** a noun in direct object position, in a non-literal expression the noun can be viewed as **selecting for** a verb.[2] We measure this property by computing the conditional probability of a verb-noun pair given the noun (CProb) which captures the degree to which the noun **selects for** a particular basic verb.

$$\begin{aligned} \text{CProb}(v,n) &\doteq P(v|n,\text{gr} = \text{dobj}) \\ &= \frac{\text{freq}(v,n|\text{gr} = \text{dobj})}{\sum_{v'} \text{freq}(v',n|\text{gr} = \text{dobj})} \end{aligned} \tag{3}$$

[2]We note that the verb selected by a noun can vary among dialects of the English language. For example, British speakers typically say *take a decision* instead of *make a decision* and *have a nap* instead of *take a nap*.

Here, we assume that children are able to keep track of the occurrences of a noun as the direct object of any verb. This measure is still a very simple one for children, since it is composed of two frequency counts.

3.1.3 Degree of Dispersion

The frequency measure formulates how many times a verb-noun compound has occurred in the data; however, it does not explain how a verb-noun compound is distributed in different parts of the data. For example, an expression can have a high frequency even though it has only occurred in one part of the data. Frequency cannot capture the difference between such expressions and an expression that occurs evenly in all parts of the data. There are other statistical measures (*e.g.*, dispersion) that capture this difference. In this section we describe a measure that uses the difference in distribution of expressions to separate literal and non-literal expressions.

We hypothesize that people's set of non-literal expressions may be very different from each other, while they use a very similar set of literal expressions when they talk. We think, because of the collacational behavior of the non-literal expressions, speakers choice depends a lot on the language they are exposed to. Consequently, we hypothesize that non-literal expressions may be used in fewer corpus divisions than the literal ones. Based on this, we calculated the degree of dispersion proposed by Gries (2008) that measures how an item is distributed in each corpus division. The following steps explain how this measure is calculated for each expression:

1. For each division i of the corpus, an expected frequency of terms used within it (EFreq) is calculated:

$$\text{EFreq}(i) \quad = \quad \frac{\text{size}(i)}{\sum_{j=1}^{n} \text{size}(j)} \tag{4}$$

   where there are n parts and $\text{size}(i)$ is size of the i^{th} part.

2. Also, an observed percentage (OPerc) for each expression (*exp*) in each part *i* is:

$$\text{OPerc}(exp, i) = \frac{\text{freq}(exp, i)}{\sum_{j=1}^{n} \text{freq}(exp, j)} \quad (5)$$

3. The Degree of dispersion (DP) for each *exp* is computed as follows:

$$\text{DP}(exp) = \frac{\sum_{j=1}^{n} |\text{OPerc}(exp, i) - \text{EFreq}(i)|}{2} \quad (6)$$

The range of DP for each expression is 0 to 1; a DP value close to 0 shows that the expression frequency in each part is related to the size of that part (as it is expected to be), while a DP value near 1 is given for an expression that occurs a lot only in few parts of the data.

In our experiments, we considered each individual file of the corpus as a division. Each file consists of some conversations between a specific child and a caretaker. As a result, we expect each non-literal expression occurs more in a single part than literal expressions.

3.2 Semantic Properties of Noun

In contrast to the noun in literal expressions, which is concrete and referential (*e.g.*, *toy* in *give the toy*), the noun in non-literal expressions appears to be more abstract and non-referential (*e.g.*, *nap* in *take a nap*). Children might use these properties of the noun to distinguish between non-literal expressions and literal expressions. We devised several measures taking into account the semantic properties of the noun that are described below.

3.2.1 Semantic Category of Noun

The noun component of a non-literal expression is expected to be an abstract noun (*e.g.*, *time* in *take a time*). We develop a measure (SCat) based on this intuition and investigate whether the semantic category of the noun component (abstract versus physical) in a verb-noun compound can be used as a clue of non-literalness.

We examined the effectiveness of the semantic category of the noun by calculating precision and recall. Precision shows to what extent abstractness of a noun predicts the non-literalness of an expression, and recall gives the proportion of non-literal expressions that actually contain an abstract noun. Precision and recall are calculated as the following conditional probabilities:

$$\text{precision} = \text{P}(\text{non-lit}(exp) = \text{YES}) \,|\, \text{semcat}(n) = \text{ABS}) \tag{7}$$

$$\text{recall} = \text{P}(\text{semcat}(n) = \text{ABS} \,|\, \text{non-lit}(exp) = \text{YES}) \tag{8}$$

Here non-lit(exp) is a function that returns YES when an expression (exp) is non-literal and NO otherwise. The semcat(n) function determines the semantic category of the noun, which is either abstract (ABS) or physical (PHY). To determine the semantic category of the noun, we look at whether it is a descendant of the ABSTRACT ENTITY or the PHYSICAL ENTITY synset in the WordNet[3] taxonomy.

The following contingency table illustrates how the conditional probabilities used in the SCat measure are calculated. Here, *tp* stands for true positive, *fp* for false positive, *tn* for true negative, and *fn* for false negative.

		non-lit(exp)	
		YES	NO
semcat(n)	ABS	*tp*	*fp*
	PHY	*fn*	*tn*

Given the above contingency table, precision and recall are calculated as follows. (We do not claim that children construct such a table to compute the probabilities. We use this notion to clarify how

[3] http://wordnet.princeton.edu/

the measure is computed.)

$$\begin{aligned} \text{precision} &= \text{P}(\text{non-lit}(exp) = \text{YES}) \,|\, \text{semcat}(n) = \text{ABS}) \\ &= \frac{tp}{tp+fp} \qquad (9) \\ \text{recall} &= \text{P}(\text{semcat}(n) = \text{ABS} \,|\, \text{non-lit}(exp) = \text{YES}) \\ &= \frac{tp}{tp+fn} \qquad (10) \end{aligned}$$

We note that this measure is in accordance with the cognitive plausibility requirement of our work. Children might be able to distinguish physical objects from the abstract ones (which do not have referents in the real world). Also, the computed conditional probabilities are simple, because they require only that children keep track of simple counts.

3.2.2 Non-referentiality

The noun component in a non-literal expression typically is non-referential, *i.e.*, it does not refer to an object in the real world. Non-referential nouns tend to appear in a particular syntactic form (Grant, 2005). For example, they are usually preceded by an indefinite determiner (such as *a, an*) or without any determiner. We develop a simple set of patterns to determine the non-referentiality of a noun in a verb-noun pair. We assume that both the determiner and the grammatical number of a noun (plural or singular) can be an indicator of the non-referential status of the noun. Below are the three patterns that we use:

$pt_{nref,0}$	$\langle \text{det} : a/an/- \;\; n_{\text{sg}} \rangle$
$pt_{nref,1}$	$\langle \text{det} : a/an/- \;\; n_{\text{sg}}/n_{\text{pl}} \rangle$
$pt_{nref,2}$	$\langle \text{det} : \text{any} \;\; n_{\text{sg}} \rangle$

where det : $a/an/-$ means that *a*, *an* or no determiner precedes the noun and det : any means that any determiner (possibly no determiner) can occur before the noun. Pattern $pt_{nref,1}$ looks at the use of the determiner regardless of whether the noun is singular or plural. Conversely, $pt_{nref,2}$ looks at the use of a singular noun, regardless of a particular determiner. Pattern $pt_{nref,0}$ combines the two

indicators by looking at the particular type of determiner with only a singular noun.

We use each pattern $pt_{nref,i}$ in a measure called $\text{NRef}_\text{i}(n)$ and calculate the probability that the noun in a verb-noun pair occurs in that pattern (see below). The numerator is the frequency of the noun in each pattern, and the denominator is the total frequency of the noun in all patterns.

$$\text{NRef}_\text{i}(n) \doteq P(pt_{nref,i}|n) = \frac{\text{freq}(n, pt_{nref,i})}{\text{freq}(n)} \tag{11}$$

3.2.3 Predicativeness

The predicative meaning of a non-literal expression is related to the noun component, which is often morphologically related to a verb (*e.g.*, *decision* in *make a decision* as the nominalized form of *decide*). To measure this, we could consider whether the noun has a related verb form and how frequent it is. To keep things simple, *i.e.*, taking into account what is realistic for a child, we instead simply count the nouns appearing as verbs without morphological change (*e.g.*, *push*). As a result, we measured the predicativeness of a noun by counting the occurrences of a noun as a verb:

$$\text{Pred}(n) \doteq \text{freq}(v'), \text{ where } \text{form}(v') = \text{form}(n) \tag{12}$$

3.3 Syntactic structure of a verb-noun pair

Non-literal expressions are known to have a fixed syntactic structure and not occur in a variety of forms (Cowie, 1981). More specifically, LVC +ABS expressions, while allowing some variation, are relatively restricted compared to LIT expressions. For example, an LVC such as *give a shout* is limited in respect to noun and determiner variations; *e.g.*, *give some shouts* and *give the shout* are not as acceptable. This is also true for ABS expressions. For example, *take a time* and *take times* are not recognized as correct variations of *take the time*. On the other hand, literal expressions are completely flexible in their syntactic structures; *e.g.*, *take an apple*, *take the apple* and *take three apples* are all acceptable structures.

In conclusion, flexibility of structure can be a distinguishing feature of literal and non-literal

expressions. Goldberg (1995) showed that children are sensitive to syntactic behavior of both words and constructions. Consequently, children might be able to recognize the syntactic structure of a verb-noun compound and use it as a cue to learn them. Below, we explain three groups of measures based on the syntactic structure of expressions. We only consider simple structures which are possible for a child to recognize.

3.3.1 Syntactic fixedness

Non-literal expressions tend to appear in fixed syntactic patterns. Researchers have measured syntactic fixedness by calculating a probability distribution over a diverse set of patterns (Fazly and Stevenson, 2007; Bannard, 2007). We use three simple patterns that draw on the non-referential and abstract status of the noun in non-literal expressions.

$pt_{pref,0}$	$\langle v \ \text{det} : a/an/- \ n_{\text{sg}} \rangle$
$pt_{pref,1}$	$\langle v \ \text{det} : a/an/- \ n_{\text{sg}}/n_{\text{pl}} \rangle$
$pt_{pref,2}$	$\langle v \ \text{det} : \text{any} \ n_{\text{sg}} \rangle$

We note that these patterns are different from the ones used in Section 3.2.2 since they also incorporate the verb. We look at the probability of each pattern for a verb-noun pair in three different measures, where i in Fixed_{i} refers to one of the three patterns above:

$$\begin{aligned} \text{Fixed}_{\text{i}}(v,n) &\doteq P(pt_{pref,i} | \, v, n, \text{gr} = \text{dobj}) \\ &= \frac{\text{freq}(v,n,pt_{pref,i} | \text{gr} = \text{dobj})}{\text{freq}(v,n | \text{gr} = \text{dobj})} \end{aligned} \tag{13}$$

3.3.2 Noun position

We hypothesize that nouns used in a non-literal combination often occur in the position of direct object of a verb, whereas nouns in literal expressions often take different positions in a sentence. The reason behind this might be that the noun in non-literal expressions is often abstract and has a non-referential status, and occurs mostly in conjunction with an appropriate verb. As a result, such nouns do not often appear in other grammatical roles. For instance, consider the literal expression

give an apple. The noun *apple* frequently takes other grammatical roles such as subject (*e.g.*, *Apples grow on trees*). On the other hand, a noun in a non-literal expression, such as *shout* or *decision*, does not occur as frequently in other positions. We formulate this property as:

$$\text{NPos}(n) \doteq \text{P}(\text{gr} = \text{dobj}|n) \quad = \quad \frac{\text{freq}(n|\text{gr} = \text{dobj})}{\text{freq}(n)} \tag{14}$$

3.3.3 Adjectival modification

Even though non-literal expressions are known to have fixed syntactic structure, nouns in LVCs are expected to take different adjectives (Akimoto and Brinton, 1999). As we described in Section 2.3, one of the possible reasons for using LVCs is their flexibility in adjectival modification (*e.g.*, *give a loud laugh* and *make an informed decision*). We study whether the adjectival modification of the noun in a verb-noun compound can be a feature of non-literal expressions. We define a pattern $pt_{adj} = \langle v\ adj\ n \rangle$ and calculate the following probability:

$$\text{AMod}(n) \doteq \text{P}(pt_{adj}|v,n,\text{gr} = \text{dobj}) \quad = \quad \frac{\text{freq}(v,n,pt_{adj}|\text{gr} = \text{dobj})}{\text{freq}(v,n|\text{gr} = \text{dobj})} \tag{15}$$

4 Experiments

In this section, we describe different experiments designed to evaluate the goodness of our measures. As input to our experiments, we use the American English section of the CHILDES database (MacWhinney, 2000), removing 16 corpora that either lack child-directed speech (CDS) or belong to a special group with a particular language use (*e.g.*, socio-economically distinguished).[4] All the data are automatically parsed with the parser of Sagae et al. (2007). Because we are interested in what is learnable from input a child is exposed to, the statistics for all experiments are extracted from CDS, except where the input type is explicitly mentioned. The size of the corpus (the CDS part) is about 600,000 utterances that contain nearly 3.2 million words (counting

[4]It has been shown that a mother's socio-economic status impacts the verbal input directed to her child, which affects the child's vocabulary acquisition (Pan et al., 2005).

punctuation).

In this work we focus on two basic verbs, *take* and *give*, because they are highly polysemous and frequently used in verb-noun combinations (Claridge, 2000). We extracted verb-noun compounds that contain these verbs from the CDS portion of the data. The final expression list that is used in the experiments includes those verb-noun compounds with a frequency of at least 5. We also restricted the data to higher-frequency verb-noun compounds, (which occurred at least 10 times), in some of the experiments. The final list of expression types is annotated by a native English speaker with four classes: LIT, ABS, LVC, and IDM. Note that we consider expression types, not tokens. If a verb-noun compound can have usages that fall into more than one class, the annotator chose the class that seemed to reflect the predominant usage.[5] Invalid expressions (due to parsing errors) and the single instance of an IDM are removed from the expression list (see the appendix on page 32). In Table 1 the number of expressions in each category and the total number of non-literal expressions (ABS+LVC) for both *freq* ≥ 5 and *freq* ≥ 10 are shown.

All expressions (*freq* ≥ 5)

V_b	Total	LIT	ABS	LVC	ABS+LVC
take	108	77	18	13	31 (29%)
give	92	75	7	10	17 (18%)
take + *give*	200	152	25	23	48 (24%)

High-frequency expressions (*freq* ≥ 10)

V_b	Total	LIT	ABS	LVC	ABS+LVC
take	57	38	8	11	19 (33%)
give	41	30	4	7	11 (27%)
take + *give*	98	68	12	18	30 (31%)

Table 1: The breakdown of the experimental expressions with freq ≥ 5, and those with freq ≥ 10.

[5]For example, the verb-noun pair *give-hand* may occur as an ABS usage (*give me a hand cleaning up*) or as a LIT usage (*give me Mr. PotatoHead's hand* or *give me your pretty hands*). In most cases of such potential ambiguity, the annotator had a clear intuition of which would be the predominant usage, since the alternative would be odd to find in CDS. In some cases, such as *give-hand*, the actual corpus usages were examined to determine the most frequent class.

4.1 Evaluating Individual Measures

We examined the performance of each individual measure in separating non-literal expressions from literal ones. Every measure assigns a score to each expression and the higher scores are indicators of the non-literalness of expressions. For each measure, expressions are sorted according to their score in descending order. As a result, we expect non-literal expressions (that have higher scores) to be placed at the beginning of the expression list. We used average precision (*AvgPrec*) to evaluate each measure. *AvgPrec* is computed as follows: Starting from the top of the expression list, for each expression its score is considered as a threshold for dividing the list into non-literal and literal parts. Then, precision is calculated for that threshold. This process is repeated until recall reaches 1. Finally, precision values for all thresholds are averaged, giving average precision.

We calculate a baseline to compare the performance of individual measures: A random value between 0 and 1 is assigned to each expression and *AvgPrec* is calculated; this process is repeated 1000 times and the average of *AvgPrec* is reported as the baseline. We note that this calculated baseline is roughly equal to the proportion of non-literal expressions in the set (the latter value appears in the final column of Table 1).

We tested the performance of each measure (except SCat) for *take* and *give* expressions separately and for all the expressions with *take* and *give*. (The SCat measure does not assign a score to each expression and cannot be evaluated using *AvgPrec*. As a result, the evaluation of this measure is discussed at the end of this section.) The results are shown in Table 2 for all the expressions with frequencies of at least 5. Nearly all measures perform better than the baseline; however, only some of them perform substantially better. (The results of those measures that perform best overall are shown in bold in the table.) Below, we look at the effectiveness of each measure.

Association of the Verb and Noun. Among the measures based on entrenchment of the verb and noun, Cooc and CProb are good indicators of non-literal expressions, which agrees with our hypothesis. Also, CProb performs somewhat better than Cooc suggesting that noun frequency is an important factor in distinguishing non-literal expressions. Degree of dispersion (DP) performs

On all expressions ($freq \geq 5$)

Measure	*take*	*give*	*take* and *give*
Baseline	$.29 \pm .04$	$.19 \pm .04$	$.24 \pm .03$
Cooc	**.53**	**.38**	**.51**
CProb	**.65**	**.47**	**.56**
DP	.23	.11	.17
$NRef_0$	**.51**	**.35**	**.42**
$NRef_1$	**.50**	**.32**	**.40**
$NRef_2$	.37	.22	.30
Pred	**.60**	**.57**	**.62**
$Fixed_0$	**.57**	**.27**	**.40**
$Fixed_1$	**.57**	**.31**	**.43**
$Fixed_2$	.31	.17	.25
Npos	.26	.14	.22
Amod	.37	.24	.31

Table 2: Performance (*AvgPrec*) of each measure on expressions with $freq \geq 5$

poorly and is even below the baseline, because the value of DP is similar for non-literal and literal expressions. That is because the number of divisions in our data (1379 divisions) is much larger than the frequency of the most frequent expression (279 for *give kiss*), which means that OPerc is 0 for many parts.

Semantic Properties of the Noun. In this group of measures, $NRef_0$ and $NRef_1$ perform reasonably well.[6] The results show that a special syntactic pattern (having *a* or *an* as determiner or no determiner) is a good sign for non-literal expressions (regardless of whether the noun is singular or plural). Also, Pred is seen to be one of the best measures, performing well not only on *take* expressions but on *give* expressions as well. This shows that the predicativeness of the noun – measured simply by how often the same form is used as a verb – is a significant distinguishing feature of non-literal expressions.

[6] $NRef_0$ and $NRef_1$ are associated with the patterns $\langle \text{det} : a/an/- \ n_{\text{sg}} \rangle$ and $\langle \text{det} : a/an/- \ n_{\text{sg}}/n_{pl} \rangle$, respectively.

Syntax of the Verb-Noun Pair. In measures related to the syntactic structure of the verb-noun pair, Fixed_0 and Fixed_1 perform much better than the baseline,[7] but Fixed_2 performs very poorly. This further confirms that determiner use is a good indicator of non-literal expressions. Additionally, NPos and AMod do not have a high *AvgPrec* score, which shows that these two properties are not very specific to non-literal measures.

Best Measures. According to the results shown in Table 2, the following measures perform better in separating non-literal and literal expressions: Cooc, CProb, $\text{NRef}_{0,1}$, Pred, and $\text{Fixed}_{0,1}$. We look at the performance of these measures on high frequency expressions (with frequencies of at least 10). For the NRef and Fixed measures, we evaluated only pt_1, which looks at all the nouns not just singular nouns. Based on the results (shown in Table 3), the performance of each measure is improved on these higher-frequency expressions (as well as the baseline). The trend is similar to the previous results for all the expressions: CProb and Pred are among the best measures, and the Fixed_1 score is much higher for *take*.

Measure	*take*	*give*	*take* and *give*
Baseline	$.33 \pm .06$	$.27 \pm .07$	$.31 \pm .05$
Cooc	.57	.41	.54
CProb	.71	.57	.64
NRef_1	.63	.47	.56
Pred	.67	.66	.68
Fixed_1	.86	.49	.66

Table 3: Performance (*AvgPrec*) of each measure on expressions with $freq \geq 10$.

Semantic Category of Noun. The measure based on the semantic category of the noun is different from the others because it does not assign a score to each expression. To evaluate this measure, we calculate precision and recall as described in Section 3.2.1 (see Table 4). We note that the taxonomy used in calculating this measure (WordNet) is very complex; therefore, it is

[7] These measures use the patterns $\langle v \text{ det} : a/an/- \; n_{\text{sg}} \rangle$ and $\langle v \text{ det} : a/an/- \; n_{\text{sg}}/n_{pl} \rangle$, which are analogous patterns to NRef_0 and NRef_1.

likely very different from children's conceptual organization. This measure would be useful if the semantic category of nouns is determined with a more plausible taxonomy. We also tried to use the MacArthur-Bates communicative developmental inventories (CDI) to determine the semantic category of nouns (Dale and Fenson, 1996).[8] However, only a small proportion of nouns in our data is included in CDI; consequently, it is not suitable for our purpose. In conclusion, despite the good performance of this measure, we do not use it in our further experiments.

SCat	*take*	*give*	*take* and *give*
precision	.58	.38	.48
recall	.68	.84	.74

Table 4: Performance of SCat on expressions with $freq \geq 5$

Conclusion. We find that very simple statistical measures based on linguistic properties of non-literal verb-noun compounds – measures which are plausibly calculable by a child – can be effective in recognizing non-literal expressions. In general, these measures perform better on the expressions composed with *take* than the expressions with *give*. A possible explanation is that the *give* expressions are more complicated, because *give* more often occurs in double object structures (in comparison to *take*). Another reason might be that the number of non-literal expressions with *give* is roughly half that of the ones with *take*. Consequently, there may not be sufficient statistical evidence available for the *give* expressions.

4.2 Clustering

By evaluating each measure individually we show its goodness as an indicator for non-literals. However, a language learner can use a combination of the available cues to recognize non-literal expressions. Consequently, we evaluated the effectiveness of a combination of the five best measures in separating non-literal expressions from literal ones using a clustering algorithm (linkage

[8] CDIs are standardized report forms that are designed "to study current behaviors and salient emergent behaviors of children that parents can recognize and track". In particular, there is a part in each form for documenting the child's production of various words divided into specific semantic categories. CDI is available at `http://www.sci.sdsu.edu/cdi/cdiwelcome.htm`.

implemented in MATLAB).[9] The linkage method creates a hierarchical agglomerative clustering tree using a dissimilarity measure (we used Euclidean distance).

The measures that are used as features in the clustering algorithm are those found to work best individually: Cooc, CProb, NRef_1, Pred, and Fixed_1. Each cluster is labeled according to the label of dominant items in the cluster. To evaluate the clusters, we used two measures: accuracy and completeness. Accuracy (*Acc*) is the proportion of expressions in a cluster that have the same label as the cluster. Completeness (*Comp*) is the fraction of all the expressions with the same label as the cluster that are placed in that cluster. *Acc* is thus similar to precision and *Comp* is similar to recall.

We performed a two-way clustering to examine the effectiveness of the measures in separating the non-literal and literal classes. We ran the clustering algorithm for both expressions with frequencies of at least 5 and expressions with frequencies of at least 10. The result are shown in Table 5.

On 200 expressions with $freq \geq 5$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	5	18	140	LIT	86%	92%
C_2	18	7	12	ABS +LVC	68%	52%

On 98 expressions with $freq \geq 10$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	1	9	64	LIT	86%	94%
C_2	17	3	4	ABS +LVC	83%	67%

Table 5: Clustering results (*Acc* and *Comp*). C_i represents Cluster *i*, Label is the label assigned to a cluster, which is the dominant label of the expression in the cluster.

In Table 5, we see that the *Acc* score for non-literal expressions is high only for the high-frequency expressions (compare C_2 in each panel of the table). We also see that literal expressions are better separated than non-literal ones since their *Comp* score is much higher (compare C_1 and C_2 for each panel of the table).

Looking closely at the number of expressions of different labels (LIT, LVC, and ABS) in each

[9] http://www.mathworks.com/access/helpdesk/help/toolbox/stats/linkage.html

cluster, it is clear that ABS expressions are more mixed with LIT expressions compared to LVC ones. Consequently, the measures are better in separating LVC from LIT than ABS from LIT. We consider the alternative goal of clustering to separate LVC expressions from the other classes (instead of separating non-literals from literals). To do so, we recalculate *Acc* and *Comp* as shown in Table 6, using LVC and LIT +ABS as the possible cluster labels. For the LVC cluster, the *Comp* score is increased (for all and higher-frequency expressions) compared to the *Comp* score calculated using LVC +ABS and LIT as the cluster labels (see Table 5). On the other hand, *Acc* is decreased. These results confirm that LVC expressions are better separated from the others than the ABS ones. ABS expressions, however, are mixed with LVC and LIT expressions.

On 200 expressions with $freq \geq 5$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	5	18	140	LIT +ABS	97%	89%
C_2	18	7	12	LVC	48%	78%

On 98 expressions with $freq \geq 10$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	1	9	64	LIT +ABS	98%	91%
C_2	17	3	4	LVC	71%	94%

Table 6: Clustering results (*Acc* and *Comp*). C_i represents Cluster *i*, Label is the label assigned to a cluster, which is the label of the majority class in the cluster.

Three-way Clustering. We also performed three-way clustering to examine the goodness of measures in dividing all expressions into ABS, LVC, and LIT classes (see Table 7). According to the results, ABS expressions did not form a separate cluster, and are mixed in the LIT and LVC clusters. In conclusion, three-way clustering partitions the previous LIT cluster into two predominantly-LIT clusters and does not perform well in building an ABS cluster. This is in accordance with our previous results: Our measures do not adequately capture properties of the ABS class.

On 200 expressions with $freq \geq 5$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	4	7	40	LIT	78%	26%
C_2	1	11	100	LIT	89%	66%
C_3	18	7	12	LVC	49%	78%

Table 7: Clustering results (*Acc* and *Comp*). C_i represents Cluster *i*, Label is the label assigned to a cluster, which is the label of the majority class in the cluster.

4.3 Error Analysis

In this section, we look over our measures again to determine which measures are good indicators of ABS expressions and which are more specific to LVC expressions. Also, we hope to find possible ways to improve the performance of our measures.

4.3.1 Evaluating Measures for ABS and LVC Expressions

We test our measures again to determine which are more specific to LVC expressions and which perform equally well for both ABS and LVC expressions. To achieve this, we examine the success of our measures first in separating only ABS expressions from LIT ones and then in separating LVC expressions from LIT ones (by calculating *AvgPrec* as described in Section 4.1). The results are shown in Table 8 and Table 9. As expected, our measures perform better for LVC expressions than ABS ones. More specifically, none of the measures is more successful for the ABS expressions. This implies that our measures are more biased towards the properties of LVC expressions. In addition, the Cooc measure score is very low for ABS expressions (close to the baseline). As a result, it might have a negative effect in our clustering result.

We re-run the clustering function described in Section 4.2 with the four best measures (eliminating the Cooc measure). The *Acc* score is slightly improved (see Table 10) compared to the *Acc* of clustering with five measures (see Table 5 on page 25). Also, the number of non-literal expressions that are clustered wrongly is decreased in the new clustering. We conclude that, while Cooc can be useful on its own in separating non-literal from literal expressions, it is not helpful in combination due to its poor performance on ABS expressions.

Measure	*take*	*give*	*take* and *give*
Baseline	$.19 \pm .04$	$.08 \pm .04$	$.14 \pm .03$
Cooc	.24	.09	.17
CProb	.46	.24	.35
$NRef_1$	.31	.11	.19
Pred	.35	.23	.33
$Fixed_1$	.41	.13	.26

Table 8: Performance (*AvgPrec*) of measures in separating ABS expressions from LIT ones ($freq \geq 5$).

Measure	*take*	*give*	*take* and *give*
Baseline	$.14 \pm .04$	$.12 \pm .04$	$.13 \pm .03$
Cooc	.43	.32	.44
CProb	.64	.39	.50
$NRef_1$	.37	.40	.32
Pred	.47	.49	.51
$Fixed_1$	.52	.28	.39

Table 9: Performance (*AvgPrec*) of measures in separating LVC expressions from LIT ones ($freq \geq 5$).

On 200 expressions with $freq \geq 5$

	LVC	ABS	LIT	Label	*Acc*	*Comp*
C_1	1	15	138	LIT	90%	91%
C_2	22	10	14	ABS +LVC	70%	67%

Table 10: Clustering results eliminating the Cooc measure. C_i represents Cluster i, Label is the label assigned to a cluster, which is the label of the majority class in the cluster.

4.3.2 C4.5 Classification

Here, instead of unsupervised clustering we use the C4.5[10] classification algorithm on our measures that builds a decision tree from a set of training data (Quinlan, 1993). A decision tree visualizes the order that features are applied in the classification task showing the importance of each feature. We do not claim that a supervised classification is similar to the way children learn MWEs. The purpose of this classification is an analysis of the contribution of each measure to the learning

[10] `http://www.rulequest.com/`

process.

We perform 10-fold cross validation on all expressions with frequency of at least 5 with the five best measures: Cooc, CProb, NRef_1, Pred, and Fixed_1.[11] The 10 decision trees (each for one fold of the run) are shown in Figure 1. Although the 10 decision trees are very different regarding the number and the order of the chosen measures, they have some interesting properties: In 7 trees out of 10, CProb or Pred (the two best measures individually) are used as the first feature in dividing the expressions. In addition, the Cooc measure is used only in 5 trees and is the last feature in 4 of the trees. Thus these results confirm the findings of our experiments on the individual measures concerning their informativity in learning.

5 Conclusion and Future Work

Our results confirm that simple statistical measures that draw on linguistic properties of non-literal expressions are useful in identifying them. The best measure for both *give* and *take* expressions is Pred, *i.e.*, the frequency of the usages of the noun as a verb. The success of this measure indicates that the predicativeness of the noun is a salient property of non-literal verb-noun combinations. In addition, the formulation of this measure is very simple, thus children could plausibly calculate it.

The second best measure for expressions with both *take* and *give* is CProb, *i.e.*, the conditional probability of a verb-noun pair given the noun. The goodness of this measure in identifying non-literal expressions suggests that the verb-noun pair in such expressions is more entrenched compared to literal ones and exhibits collocational behavior. However, collocational behavior alone is not a very good indicator of non-literal expressions; the CProb measure consistently outperforms Cooc (which only quantifies the entrenchment of the verb-noun pair). The key difference between these two measures is that in CProb, we also measure the degree that the noun **selects for** the appropriate verb.

[11] The program takes two parameters: -m and -c. The -m parameter is the minimum number of items required in each branch of decision tree. The -c option is a confidence factor in pruning. We get the best accuracy (81%) when $m = 8$ and $c = 15$.

```
NRef <= 0.822969 :
|   Fixed <= 0.4 : l (130/9)
|   Fixed > 0.4 :
|   |   Cooc <= 0.0286738 : l (13/4)
|   |   Cooc > 0.0286738 : a (10/1)
NRef > 0.822969 :
|   Pred <= 0.00558952 : l (10/4)
|   Pred > 0.00558952 : a (17)
```

(a) tree1

```
Pred <= 0.00925764 :
|   CProb > 0.387097 : a (9/1)
|   CProb <= 0.387097 :
|   |   Fixed <= 0.428571 : l (121/7)
|   |   Fixed > 0.428571 :
|   |   |   CProb <= 0.0760234 : l (13/2)
|   |   |   CProb > 0.0760234 : a (8/3)
Pred > 0.00925764 :
|   NRef <= 0.822969 : l (12/4)
|   NRef > 0.822969 : a (17)
```

(b) tree2

```
CProb <= 0.0913979 : l (114/7)
CProb > 0.0913979 :
|   CProb > 0.294118 : a (19/1)
|   CProb <= 0.294118 :
|   |   Fixed <= 0.428571 : l (34/8)
|   |   Fixed > 0.428571 : a (13/3)
```

(c) tree3

```
CProb <= 0.0913979 : l (120/8)
CProb > 0.0913979 :
|   Fixed <= 0.111111 : l (9)
|   Fixed > 0.111111 :
|   |   Fixed > 0.4 : a (21/2)
|   |   Fixed <= 0.4 :
|   |   |   NRef > 0.811594 : a (9)
|   |   |   NRef <= 0.811594 :
|   |   |   |   Cooc <= 0.046595 : l (10)
|   |   |   |   Cooc > 0.046595 : a (11/4)
```

(d) tree4

```
CProb <= 0.0947867 : l (119/8)
CProb > 0.0947867 :
|   Fixed <= 0.111111 : l (8)
|   Fixed > 0.111111 :
|   |   Pred > 0.0450655 : a (12)
|   |   Pred <= 0.0450655 :
|   |   |   CProb > 0.5 : a (9)
|   |   |   CProb <= 0.5 :
|   |   |   |   Fixed > 0.5 : a (8/2)
|   |   |   |   Fixed <= 0.5 :
|   |   |   |   |   Cooc <= 0.0501792 : l (16/3)
|   |   |   |   |   Cooc > 0.0501792 : a (8/3)
```

(e) tree5

```
CProb <= 0.0913979 : l (114/7)
CProb > 0.0913979 :
|   NRef > 0.811594 : a (21/2)
|   NRef <= 0.811594 :
|   |   Fixed > 0.4 : a (11/2)
|   |   Fixed <= 0.4 :
|   |   |   NRef <= 0.392265 : l (11)
|   |   |   NRef > 0.392265 :
|   |   |   |   Cooc <= 0.0430108 : l (12/2)
|   |   |   |   Cooc > 0.0430108 : a (11/5)
```

(f) tree6

```
Pred <= 0.00925764 :
|   CProb <= 0.387097 : l (143/13)
|   CProb > 0.387097 : a (10/1)
Pred > 0.00925764 :
|   NRef <= 0.822969 : l (11/5)
|   NRef > 0.822969 : a (16)
```

(g) tree7

```
Pred <= 0.00925764 :
|   CProb <= 0.0913979 : l (112/5)
|   CProb > 0.0913979 :
|   |   Fixed <= 0.4 : l (30/8)
|   |   Fixed > 0.4 : a (14/2)
Pred > 0.00925764 :
|   Cooc <= 0.0286738 : l (9/4)
|   Cooc > 0.0286738 : a (15)
```

(h) tree8

```
NRef > 0.805668 : a (28/6)
NRef <= 0.805668 :
|   Fixed <= 0.4 : l (134/10)
|   Fixed > 0.4 :
|   |   CProb <= 0.0707071 : l (8/3)
|   |   CProb > 0.0707071 : a (10/1)
```

(i) tree9

```
NRef <= 0.805668 :
|   CProb <= 0.0913979 : l (108/5)
|   CProb > 0.0913979 :
|   |   CProb <= 0.28169 : l (34/9)
|   |   CProb > 0.28169 : a (8/1)
NRef > 0.805668 :
|   Pred <= 0.00558952 : l (14/6)
|   Pred > 0.00558952 : a (16)
```

(j) tree10

Figure 1: Decision trees

In addition, based on the result of evaluating the measures individually and in clustering, our measures are generally better for high-frequency expressions. In particular, the performance of the NRef_1 and Fixed_1 measures that rely on syntactic patterns improves notably for high-frequency expressions. Consequently, more frequent expression are easier to learn. However, the two best measures (Pred and CProb) perform well on both expressions with frequency of at least 5 and high-frequency expressions, suggesting that the children might be able to learn MWEs even with very little data.

Our results also show that the performance of our measures is better for *take* expressions compared to *give* expressions (even in high-frequency expressions). This suggests that children might find the expressions containing *give* harder to learn than the ones with *take*. One possible explanation is that the MWEs with *give* are syntactically more complex.

Moreover, the measures can better separate the light verb constructions than the abstract expressions. This is because some measures draw on specific properties of light verb constructions. Also, the distinguishing properties of the abstract class are not well explored in the literature.

In conclusion, we show that the statistical evidence available in the input children receive can be used in identifying non-literal expressions. We also devised very simple measures that are reasonable for young children. In future, we would like to show how these measure can be embedded in a model of word learning. This model would show how children can recognize and learn the meaning of multiword expressions.

Appendix

The expressions are annotated by a native speaker of English following the annotation guidelines in Fazly (2007).

Table 11: **The list of *give* expressions in class ABS**

Expression	**Frequency**	**Expression**	**Frequency**
give bottle	25	give love	9
give medicine	22	give second	8
give name	17	give rest	7
give shot	12	give rest	7

Table 12: **The list of *give* expressions in class LVC**

Expression	**Frequency**	**Expression**	**Frequency**
give kiss	279	give drink	12
give hug	89	give push	11
give ride	55	give hint	9
give bath	32	give try	7
give bite	20	give spank	5

Table 13: **The list of *give* expressions in class LIT**

Expression	**Frequency**	**Expression**	**Frequency**
give one	152	give bread	8
give hand	59	give who	7
give money	58	give two	7
give ball	57	give sugar	7
give piece	36	give lollipop	7
give cookie	28	give icecream	7
give juice	27	give horse	7
give milk	25	give comb	7
give dog	22	give carrot	7
give bone	20	give tea	6
give change	19	give present	6
give doll	17	give penny	6
give cup	17	give nickel	6
give book	17	give letter	6
give baby	17	give leg	6
give pencil	16	give finger	6
give food	16	give crayon	6

Continued on next page

Table 13 – continued from previous page

Expression	Frequency	Expression	Frequency
give foot	15	give cracker	6
give coffee	15	give cent	6
give spoon	14	give car	6
give shampoo	14	give boy	6
give paper	14	give blanket	6
give dollar	13	give bear	6
give bit	13	give treat	5
give box	12	give tape	5
give thing	11	give pipe	5
give quarter	11	give pet	5
give pen	11	give person	5
give block	11	give part	5
give toy	10	give orange	5
give lot	9	give monkey	5
give fish	9	give honey	5
give candy	9	give half	5
give water	8	give egg	5
give lunch	8	give cream	5
give gas	8	give brush	5

Continued on next page

Table 13 – concluded from previous page

Expression	Frequency	Expression	Frequency
give cheese	8	give bill	5
give breakfast	8	give bill	5

Table 14: **The list of *take* expressions in class ABS**

Expression	Frequency	Expression	Frequency
take picture	108	take medicine	9
take car	41	take bus	9
take time	28	take train	7
take turn	24	take minute	7
take step	15	take cover	7
take temperature	14	take side	6
take trip	13	take lesson	6
take truck	10	take drive	6
take while	9	take practice	5

Table 15: **The list of *take* expressions in class LVC**

Expression	Frequency	Expression	Frequency
take nap	213	take look	23
take bath	143	take rest	21
take bite	80	take shower	18
take care	75	take sip	11
take drink	56	take guess	7
take ride	38	take sleep	5
take walk	36	take sleep	5

Table 16: **The list of *take* expressions in class LIT**

Expression	Frequency	Expression	Frequency
take one	150	take person	9
take shoe	39	take lid	9
take book	39	take foot	9
take thing	38	take two	8
take wheel	35	take tape	8
take clothes	28	take duck	8
take piece	27	take color	8
take finger	26	take bottle	8
take sock	25	take bag	8
take home	25	take stuff	7
take hand	25	take movie	7
take paper	24	take eye	7
take top	23	take egg	7
take coat	22	take dress	7
take tire	20	take diaper	7
take toy	18	take cup	7
take dog	18	take crayon	7

Continued on next page

Table 16 – continued from previous page

Expression	Frequency	Expression	Frequency
take shirt	17	take child	7
take pants	17	take cat	7
take ball	17	take airplane	7
take box	16	take water	6
take money	15	take stick	6
take doll	15	take skin	6
take bib	15	take pocketbook	6
take nose	14	take microphone	6
take hair	14	take lady	6
take glass	14	take chair	6
take pencil	13	take suitcase	5
take block	13	take puzzle	5
take baby	13	take letter	5
take man	12	take key	5
take sweater	11	take food	5
take string	11	take cookie	5
take spoon	11	take cheese	5
take head	11	take bump	5
take somewhere	10	take blanket	5

Continued on next page

Table 16 – concluded from previous page

Expression	Frequency	Expression	Frequency
take part	10	take bit	5
take hat	10	take arm	5
take phone	9	take arm	5

References

M. Akimoto and L. J. Brinton, editors. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Studies in Language Companion Series 47. John Benjamins, 1999.

A. Alishahi and S. Stevenson. A computational model of early argument structure acquisition. *Cognitive Science: A Multidisciplinary Journal*, 32(5):789–834, 2008.

C. Bannard. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Multi Word Expression'07: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics, 2007.

K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum, 1991.

C. Claridge. *Multi-Word Verbs in Early Modern English*. Language and Computers 32. Rodopi, New York, 2000.

A. Clark. Unsupervised induction of stochastic context free grammars with distributional clustering. In *Proceedings of Conference on Computational Natural Language Learning*, pages 105–112, 2001.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

A. P. Cowie. The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, II(3):223–235, 1981.

A. P. Cowie, R. Mackin, and I. McCaig. *Oxford dictionary of current idiomatic English*. Oxford University Press, 1975.

P. S. Dale and L. Fenson. Lexical development norms for young children. *Behavior Research Methods, Instrumentation, and Computers*, 28:125–127, 1996.

M. Dras and M. Johnson. Death and lightness: Using a demographic model to find support verbs. In *In Proceedings of the Fifth International Conference on the Cognitive Science of Natural Language Processing*, pages 165–172, 1996.

S. Evert. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, 2008. Article 58.

S. Evert, U. Heid, and K. Spranger. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th Int'l Conference on Language Resources and Evaluation*, pages 907–910, 2004.

A. Fazly. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. PhD in Computer Science, University of Toronto, 2007.

A. Fazly and S. Stevenson. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Multi Word Expression'07: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 9–16. Association for Computational Linguistics, 2007.

A. Fazly, A. Alishahi, and S. Stevenson. A probabilistic computational model of cross-situational word learning. *To appear in Cognitive Science*.

A. Fazly, S. Stevenson, and R. North. Automatically learning semantic knowledge about multiword predicates. *Journal of Language Resources and Evaluation*, 41(1):61–89, 2007.

A. Fazly, P. Cook, and S. Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103, 2009.

M. Frank, N. Goodman, and J. B. Tenenbaum. A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*, 2007.

A. E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press, 1995.

L. E. Grant. Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics*, 10(4):429–451, 2005.

G. Grefenstette and S. Teufel. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 98–103, 1995.

S. T. Gries. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437, 2008.

D. Lin. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics, 1999.

B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*, volume 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates, 3rd edition, 2000.

B. A. Pan, M. L. Rowe, J. D. Singer, and C. E. Snow. Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, 76(4):763–782, July 2005.

A. H. Pollard, F. Yusuf, and G. N. Pollard. *Demographic techniques*. Pergamon Press, Sydney, 2nd edition, 1981.

J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the ACL'07 Workshop on Cognitive Aspects of Computational Language Acquisition*, 2007.

W. Sakas and J. D. Fodor. The structural triggers learner. In S. Bertolo, editor, *Language Acquistion and Learnability*, pages 172–233. Cambridge University Press, Cambridge, UK, 2001.

F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

J. B. Tenenbaum and F. Xu. Word learning as bayesian inference. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 517–522. Erlbaum, 2000.

M. Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard Univ. Press, 2003.