Unsupervised Type and Token Identification of Idiomatic Expressions

Afsaneh Fazly*
University of Toronto

Paul Cook**
University of Toronto

Suzanne Stevenson†
University of Toronto

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Idiomatic expressions are plentiful in everyday language, yet they remain mysterious, as it is not clear exactly how people learn and understand them. They are of special interest to linguists, psycholinguists, and lexicographers, mainly because of their syntactic and semantic idiosyncrasies as well as their unclear lexical status. Despite a great deal of research on the properties of idioms in the linguistics literature, there is not much agreement on which properties are characteristic of these expressions. Because of their peculiarities, idiomatic expressions have mostly been overlooked by researchers in computational linguistics. In this article, we look into the usefulness of some of the identified linguistic properties of idioms for their automatic recognition. Specifically, we develop statistical measures that each model a specific property of idiomatic expressions by looking at their actual usage patterns in text. We use these statistical measures in a type-based classification task where we automatically separate idiomatic expressions (expressions with a possible idiomatic interpretation) from similar-on-the-surface literal phrases (for which no idiomatic interpretation is possible). In addition, we use some of the measures in a token identification task where we distinguish idiomatic and literal usages of potentially idiomatic expressions in context.

1. Introduction

Idioms form a heterogeneous class, with prototypical examples such as *by and large*, *kick the bucket*, and *let the cat out of the bag*. It is hard to find a single agreed-upon definition that covers all members of this class (Glucksberg 1993; Cacciari 1993; Nunberg, Sag, and Wasow 1994), but they are often defined as sequences of words involving some degree of semantic idiosyncrasy or non-compositionality. That is, an idiom has a different

* Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4, Canada. E-mail: afsaneh@cs.toronto.edu.
** Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4, Canada. E-mail: pcook@cs.toronto.edu.
† Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4, Canada. E-mail: suzanne@cs.toronto.edu.

Submission received: 12 September 2007; revised submission received: 29 February 2008; accepted for publication: 6 May 2008.

© 2009 Association for Computational Linguistics

meaning from the simple composition of the meaning of its component words. Idioms are widely and creatively used by speakers of a language to express ideas cleverly, economically, or implicitly, and thus appear in all languages and in all text genres (Sag et al. 2002). Many expressions acquire an idiomatic meaning over time (Cacciari 1993); consequently, new idioms come into existence on a daily basis (Cowie, Mackin, and McCaig 1983; Seaton and Macaulay 2002). Automatic tools are therefore necessary for assisting lexicographers in keeping lexical resources up to date, as well as for creating and extending computational lexicons for use in natural language processing (NLP) systems.

Though completely frozen idioms, such as *by and large*, can be represented as words with spaces (Sag et al. 2002), most idioms are syntactically well-formed phrases that allow some variability in expression, such as *shoot the breeze* and *hold fire* (Gibbs and Nayak 1989; d'Arcais 1993; Fellbaum 2007). Such idioms allow a varying degree of morphosyntactic flexibility—for example, *held fire* and *hold one's fire* allow for an idiomatic reading, whereas typically only a literal interpretation is available for *fire was held* and *held fires*. Clearly, a words-with-spaces approach does not work for phrasal idioms. Hence, in addition to requiring NLP tools for recognizing idiomatic expressions (types) to include in a lexicon, methods for determining the allowable and preferred usages (a.k.a. canonical forms) of such expressions are also needed. Moreover, in many situations, an NLP system will need to distinguish a usage (token) of a potentially idiomatic expression as either idiomatic or literal in order to handle a given sequence of words appropriately. For example, a machine translation system must translate *held fire* differently in *The army held their fire* and *The worshippers held the fire up to the idol*.

Previous studies focusing on the automatic identification of idiom types have often recognized the importance of drawing on their linguistic properties, such as their semantic idiosyncrasy or their restricted flexibility, pointed out earlier. Some researchers have relied on a manual encoding of idiom-specific knowledge in a lexicon (Copestake et al. 2002; Odijk 2004; Villavicencio et al. 2004), whereas others have presented approaches for the automatic acquisition of more general (hence less distinctive) knowledge from corpora (Smadja 1993; McCarthy, Keller, and Carroll 2003). Recent work that looks into the acquisition of the distinctive properties of idioms has been limited, both in scope and in the evaluation of the methods proposed (Lin 1999; Evert, Heid, and Spranger 2004). Our goal is to develop unsupervised means for the automatic acquisition of lexical, syntactic, and semantic knowledge about a broadly documented class of idiomatic expressions.

Specifically, we focus on a cross-linguistically prominent class of phrasal idioms which are commonly and productively formed from the combination of a frequent verb and a noun in its direct object position (Cowie, Mackin, and McCaig 1983; Nunberg, Sag, and Wasow 1994; Fellbaum 2002), for example, *shoot the breeze*, *make a face*, and *push one's luck*. We refer to these as verb+noun idiomatic combinations or VNICs.[1] We present a comprehensive analysis of the distinctive linguistic properties of phrasal idioms, including VNICs (Section 2), and propose statistical measures that capture each property (Section 3). We provide a multi-faceted evaluation of the measures (Section 4), showing their effectiveness in the recognition of idiomatic expressions (types)—that is, separating them from similar-on-the-surface literal phrases—as well as their superiority to existing state-of-the-art techniques. Drawing on these statistical measures, we also propose an unsupervised method for the automatic acquisition of an idiom's canonical

1 We use the abbreviation VNIC and the term **expression** to refer to a verb+noun type with a potential idiomatic meaning. We use the terms **instance** and **usage** to refer to a token occurrence of an expression.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

forms (e.g., *shoot the breeze* as opposed to *shoot a breeze*), and show that it can successfully accomplish the task (Section 5).

It is possible for a single VNIC to have both idiomatic and non-idiomatic (literal) meanings. For example, *make a face* is ambiguous between an idiom, as in *The little girl made a funny face at her mother*, and a literal combination, as in *She made a face on the snowman using a carrot and two buttons*. Despite the common perception that phrases that can be idioms are mainly used in their idiomatic sense, our analysis of 60 idioms has shown otherwise. We found that close to half of these also have a clear literal meaning; and of those with a literal meaning, on average around 40% of their usages are literal. Distinguishing token phrases as idiomatic or literal combinations of words is thus essential for NLP tasks, such as semantic parsing and machine translation, which require the identification of multiword semantic units.

Most recent studies focusing on the identification of idiomatic and non-idiomatic tokens either assume the existence of manually annotated data for a supervised classification (Patrick and Fletcher 2005; Katz and Giesbrecht 2006), or rely on manually encoded linguistic knowledge about idioms (Uchiyama, Baldwin, and Ishizaki 2005; Hashimoto, Sato, and Utsuro 2006), or even ignore the specific properties of non-literal language and rely mainly on general purpose methods for the task (Birke and Sarkar 2006). We propose unsupervised methods that rely on automatically acquired knowledge about idiom types to identify their token occurrences as idiomatic or literal (Section 6). More specifically, we explore the hypothesis that the type-based knowledge we automatically acquire about an idiomatic expression can be used to determine whether an instance of the expression is used literally or idiomatically (token-based knowledge). Our experimental results show that the performance of the token-based idiom identification methods proposed here is comparable to that of existing supervised techniques (Section 7).

2. Idiomaticity, Semantic Analyzability, and Flexibility

Although syntactically well-formed, phrasal idioms (including VNICs) involve a certain degree of semantic idiosyncrasy. This means that phrasal idioms are to some extent nontransparent; that is, even knowing the meaning of the individual component words, the meaning of the idiom is hard to determine without special context or previous exposure. There is much evidence in the linguistics literature that idiomatic combinations also have idiosyncratic lexical and syntactic behavior. Here, we first define semantic analyzability and elaborate on its relation to semantic idiosyncrasy or idiomaticity. We then expound on the lexical and syntactic behavior of VNICs, pointing out a suggestive relation between the degree of idiomaticity of a VNIC and the degree of its lexicosyntactic flexibility.

2.1 Semantic Analyzability

Idioms have been traditionally believed to be completely non-compositional (Fraser 1970; Katz 1973). This means that unlike compositional combinations, the meaning of an idiom cannot be solely predicted from the meaning of its parts. Nonetheless, many linguists and psycholinguists argue against such a view, providing evidence from idioms that show some degree of semantic compositionality (Nunberg, Sag, and Wasow 1994; Gibbs 1995). The alternative view suggests that many idioms in fact do

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

have internal semantic structure, while recognizing that they are not compositional in a simplistic or traditional sense. To explain the semantic behavior of idioms, researchers who take this alternative view thus use new terms such as **semantic decomposability** and/or **semantic analyzability** in place of compositionality.

To say that an idiom is semantically analyzable to some extent means that the constituents contribute some sort of independent meaning—not necessarily their literal semantics—to the overall idiomatic interpretation. Generally, the more semantically analyzable an idiom is, the easier it is to map the idiom constituents onto their corresponding idiomatic referents. In other words, the more semantically analyzable an idiom is, the easier it is to make predictions about the idiomatic meaning from the meaning of the idiom parts. Semantic analyzability is thus inversely related to semantic idiosyncrasy.

Many linguists and psycholinguists conclude that idioms clearly form a heterogeneous class, not all of them being truly non-compositional or unanalyzable (Abeillé 1995; Moon 1998; Grant 2005). Rather, semantic analyzability in idioms is a matter of degree. For example, the meaning of *shoot the breeze* ("to chat idly"), a highly idiomatic expression, has nothing to do with either *shoot* or *breeze*. A less idiomatic expression, such as *spill the beans* ("to reveal a secret"), may be analyzed as *spill* metaphorically corresponding to "reveal" and *beans* referring to "secret(s)." An idiom such as *pop the question* is even less idiomatic because the relations between the idiom parts and their idiomatic referents are more directly established, namely, *pop* corresponds to "suddenly ask" and *question* refers to "marriage proposal." As we will explain in the following section, there is evidence that the difference in the degree of semantic analyzability of idiomatic expressions is also reflected in their lexical and syntactic behavior.

2.2 Lexical and Syntactic Flexibility

Most idioms are known to be lexically fixed, meaning that the substitution of a near synonym (or a closely related word) for a constituent part does not preserve the idiomatic meaning of the expression. For example, neither *shoot the wind* nor *hit the breeze* are valid variations of the idiom *shoot the breeze*. Similarly, *spill the beans* has an idiomatic meaning, while *spill the peas* and *spread the beans* have only literal interpretations. There are, however, idiomatic expressions that have one (or more) lexical variants. For example, *blow one's own trumpet* and *toot one's own horn* have the same idiomatic interpretation (Cowie, Mackin, and McCaig 1983); also *keep one's cool* and *lose one's cool* have closely related meanings (Nunberg, Sag, and Wasow 1994). Nonetheless, it is not the norm for idioms to have lexical variants; when they do, there are usually unpredictable restrictions on the substitutions they allow.

Idiomatic combinations are also syntactically distinct from compositional combinations. Many VNICs cannot undergo syntactic variations and at the same time retain their idiomatic interpretations. It is important, however, to note that VNICs differ with respect to the extent to which they can tolerate syntactic operations, that is, the degree of syntactic flexibility they exhibit. Some are syntactically inflexible for the most part, whereas others are more versatile, as illustrated in the sentences in Examples (1) and (2):

1. (a) Sam and Azin shot the breeze.
   (b) ?? Sam and Azin shot a breeze.
   (c) ?? Sam and Azin shot the breezes.
   (d) ?? Sam and Azin shot the casual breeze.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

(e) ?? The breeze was shot by Sam and Azin.
(f) ?? The breeze that Sam and Azin shot was quite refreshing.
(g) ?? Which breeze did Sam and Azin shoot?

2. (a) Azin spilled the beans.
(b) ? Azin spilled some beans.
(c) ?? Azin spilled the bean.
(d) Azin spilled the Enron beans.
(e) The beans were spilled by Azin.
(f) The beans that Azin spilled caused Sam a lot of trouble.
(g) Which beans did Azin spill?

Linguists have often explained the lexical and syntactic flexibility of idiomatic combinations in terms of their semantic analyzability (Fellbaum 1993; Gibbs 1993; Glucksberg 1993; Nunberg, Sag, and Wasow 1994; Schenk 1995). The common belief is that because the constituents of a semantically analyzable idiom can be mapped onto their corresponding referents in the idiomatic interpretation, analyzable (less idiomatic) expressions are often more open to lexical substitution and syntactic variation. Psycholinguistic studies also support this hypothesis: Gibbs and Nayak (1989) and Gibbs et al. (1989), through a series of psychological experiments, demonstrate that there is variation in the degree of lexicosyntactic flexibility of idiomatic combinations. (Both studies narrow their focus to verb phrase idiomatic combinations, mainly of the form verb+noun.) Moreover, their findings provide evidence that the lexical and syntactic flexibility of VNICs is not arbitrary, but rather correlates with the semantic analyzability of these idioms as perceived by the speakers participating in the experiments.

Corpus-based studies such as those by Moon (1998), Riehemann (2001), and Grant (2005) conclude that idioms are not as fixed as most have assumed. These claims are often based on observing certain idiomatic combinations in a form other than their so-called canonical forms. For example, Moon mentions that she has observed both *kick the pail* and *kick the can* as variations of *kick the bucket*. Also, Grant finds evidence of variations such as *eat one's heart (out)* and *eat one's hearts (out)* in the BNC. Riehemann concludes that in contrast to non-idiomatic combinations of words, "idioms have a strongly preferred canonical form, but at the same time the occurrence of lexical and syntactic variations of idioms is too common to be ignored" (page 67). Our understanding of such findings is that idiomatic combinations are not inherently frozen and that it is possible for them to appear in forms other than their agreed-upon canonical forms. However, it is important to note that most such observed variations are constrained, often with unpredictable restrictions.

We are well aware that semantic analyzability is neither a necessary nor a sufficient condition for an idiomatic combination to be lexically or syntactically flexible. Other factors, such as communicative intentions and pragmatic constraints, can motivate a speaker to use a variant in place of a canonical form (Glucksberg 1993). For example, journalism is well known for manipulating idiomatic expressions for humor or cleverness (Grant 2005). The age and the degree of familiarity of an idiom have also been shown to be important factors that affect its flexibility (Gibbs and Nayak 1989). Nonetheless, linguists often use observations about lexical and syntactic flexibility of VNICs in order to make judgments about their degree of idiomaticity (Kytö 1999; Tanabe 1999). We thus conclude that lexicosyntactic behavior of a VNIC, although affected by historical and pragmatic factors, can be at least partially explained in terms of semantic analyzability or idiomaticity.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

3. Automatic Acquisition of Type-Based Knowledge about VNICs

We use the observed connection between idiomaticity and (in)flexibility to devise statistical measures for automatically distinguishing idiomatic verb+noun combinations (types) from literal phrases. More specifically, we aim to identify verb–noun pairs such as ⟨*keep*, *word*⟩ as having an associated idiomatic expression (*keep one's word*), and also distinguish these from verb–noun pairs such as ⟨*keep*, *fish*⟩ which do not have an idiomatic interpretation. Although VNICs vary in their degree of flexibility (cf. Examples (1) and (2)), on the whole they contrast with fully compositional phrases, which are more lexically productive and appear in a wider range of syntactic forms. We thus propose to use the degree of lexical and syntactic flexibility of a given verb+noun combination to determine the level of idiomaticity of the expression.

Note that our assumption here is in line with corpus-linguistic studies on idioms: we do not claim that it is inherently impossible for VNICs to undergo lexical substitution or syntactic variation. In fact, for each given idiomatic combination, it may well be possible to find a specific situation in which a lexical or a syntactic variant of the canonical form is perfectly plausible. However, the main point of the assumption here is that VNICs are more likely to appear in fixed forms (known as their canonical forms), more so than non-idiomatic phrases. Therefore, the overall distribution of a VNIC in different lexical and syntactic forms is expected to be notably different from the corresponding distribution of a typical verb+noun combination.

The following subsections describe our proposed statistical measures for idiomaticity, which quantify the degree of lexical, syntactic, and overall fixedness of a given verb+noun combination (represented as a verb–noun pair).

3.1 Measuring Lexical Fixedness

A VNIC is lexically fixed if the replacement of any of its constituents by a semantically (and syntactically) similar word does not generally result in another VNIC, but in an invalid or a literal expression. One way of measuring lexical fixedness of a given verb+noun combination is thus to examine the idiomaticity of its variants, that is, expressions generated by replacing one of the constituents by a similar word. This approach has two main challenges: (i) it requires prior knowledge about the idiomaticity of expressions (which is what we are developing our measure to determine); (ii) it can only measure the lexical fixedness of idiomatic combinations, and so could not apply to literal combinations. We thus interpret this property statistically in the following way: We expect a lexically fixed verb+noun combination to appear much more frequently than its variants in general.

Specifically, we examine the strength of association between the verb and the noun constituent of a combination (the target expression or its lexical variants) as an indirect cue to its idiomaticity, an approach inspired by Lin (1999). We use the automatically built thesaurus of Lin (1998) to find words similar to each constituent, in order to automatically generate variants.[2] Variants are generated by replacing either

2 We also replicated our experiments with an automatically built thesaurus created from the British National Corpus (BNC) in a similar fashion, and kindly provided to us by Diana McCarthy. Results were similar, hence we do not report them here.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

the noun or the verb constituent of a pair with a semantically (and syntactically) similar word.[3]

Examples of automatically generated variants for the pair ⟨*spill*, *bean*⟩ are ⟨*pour*, *bean*⟩, ⟨*stream*, *bean*⟩, ⟨*spill*, *corn*⟩, and ⟨*spill*, *rice*⟩.

Let $\mathcal{S}_{sim}(v) = \{v_i \mid 1 \leq i \leq K_v\}$ be the set of the K_v most similar verbs to the verb v of the target pair $\langle v, n\rangle$, and $\mathcal{S}_{sim}(n) = \{n_j \mid 1 \leq j \leq K_n\}$ be the set of the K_n most similar nouns to the noun n (according to Lin's thesaurus). The set of variants for the target pair is thus:

$$\mathcal{S}_{sim}(v, n) = \{\langle v_i, n\rangle | \, 1 \leq i \leq K_v\} \cup \{\langle v, n_j\rangle | \, 1 \leq j \leq K_n\}.$$

We calculate the association strength for the target pair and for each of its variants using an information-theoretic measure called pointwise mutual information or PMI (Church et al. 1991):

$$\begin{aligned} \text{PMI}(v_r, n_t) &= \log \frac{P(v_r, n_t)}{P(v_r)\,P(n_t)} \\ &= \log \frac{\text{N}_{v+n}\, f(v_r, n_t)}{f(v_r, *)\, f(*, n_t)} \end{aligned} \tag{1}$$

where $\langle v_r, n_t\rangle \in \{\langle v, n\rangle\} \cup \mathcal{S}_{sim}(v, n)$; N_{v+n} is the total number of verb–object pairs in the corpus; $f(v_r, n_t)$ is the frequency of v_r and n_t co-occurring as a verb–object pair; $f(v_r, *)$ is the total frequency of the target (transitive) verb with any noun as its direct object; and $f(*, n_t)$ is the total frequency of the noun n_t in the direct object position of any verb in the corpus.

In his work, Lin (1999) assumes that a target expression is non-compositional if and only if its PMI value is significantly different from that of all the variants. Instead, we propose a novel technique that brings together the association strengths (PMI values) of the target and the variant expressions into a single measure reflecting the degree of lexical fixedness for the target pair. We assume that the target pair is lexically fixed to the extent that its PMI deviates from the average PMI of its variants. By our measure, the target pair is considered lexically fixed (i.e., is given a high fixedness score) only if the difference between its PMI value and that of most of its variants—not necessarily all, as in the method of Lin (1999)—is high.[4] Our measure calculates this deviation, normalized using the sample's standard deviation:

$$\text{Fixedness}_{\text{lex}}(v, n) \doteq \frac{\text{PMI}(v, n) - \overline{\text{PMI}}}{s} \tag{2}$$

3 In an early version of this work (Fazly and Stevenson 2006), only the noun constituent was varied because we expected replacing the verb constituent with a related verb to be more likely to yield another VNIC, as in *keep/lose one's cool*, *give/get the bird*, *crack/break the ice* (Nunberg, Sag, and Wasow 1994; Grant 2005). Later experiments on the development data showed that variants generated by replacing both constituents, one at a time, produce better results.

4 This way, even if an idiom has a few frequently used variants (e.g., *break the ice* and *crack the ice*), it may still be assigned a high fixedness score if most other variants are uncommon. Note also that it is possible that some variants of a given idiom are frequently used literal expressions (e.g., *make biscuit* for *take biscuit*). It is thus important to use a flexible formulation that relies on the collective evidence (e.g., average PMI) and hence is less sensitive to individual cases.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

where $\overline{\text{PMI}}$ is the mean and s the standard deviation of the following sample:

$$\left\{\text{PMI}(v_r, n_t) \mid \langle v_r, n_t\rangle \in \{\langle v, n\rangle\} \cup \mathcal{S}_{sim(v,n)}\right\}$$

PMI can be negative, zero, or positive; thus $\text{Fixedness}_{\text{lex}}(v, n) \in [-\infty, +\infty]$, where high positive values indicate higher degrees of lexical fixedness.

3.2 Measuring Syntactic Fixedness

Compared to literal (non-idiomatic) verb+noun combinations, VNICs are expected to appear in more restricted syntactic forms. To quantify the syntactic fixedness of a target verb–noun pair, we thus need to: (i) identify relevant syntactic patterns, namely, those that help distinguish VNICs from literal verb+noun combinations; and (ii) translate the frequency distribution of the target pair in the identified patterns into a measure of syntactic fixedness.

3.2.1 Identifying Relevant Patterns. Determining a unique set of syntactic patterns appropriate for the recognition of all idiomatic combinations is difficult indeed: Exactly which forms an idiomatic combination can occur in is not entirely predictable (Sag et al. 2002). Nonetheless, there are hypotheses about the difference in behavior of VNICs and literal verb+noun combinations with respect to particular syntactic variations (Nunberg, Sag, and Wasow 1994). Linguists note that semantic analyzability of VNICs is related to the referential status of the noun constituent (i.e., the process of idiomatization of a verb+noun combination is believed to be accompanied by a change from concreteness to abstractness for the noun). The referential status of the noun is in turn assumed to be related to the participation of the combination in certain morpho-syntactic forms. In what follows, we describe three types of syntactic variation that are assumed to be mostly tolerated by literal combinations, but less tolerated by many VNICs.

Passivization. There is much evidence in the linguistics literature that VNICs often do not undergo passivization. Linguists mainly attribute this to the fact that in most cases, only referential nouns appear as the surface subject of a passive construction (Gibbs and Nayak 1989). Due to the non-referential status of the noun constituent in most VNICs, we expect that they do not undergo passivization as often as literal verb+noun combinations do. Another explanation for this assumption is that passives are mainly used to put focus on the object of a clause or sentence. For most VNICs, no such communicative purpose can be served by topicalizing the noun constituent through passivization (Jackendoff 1997). The passive construction is thus considered as one of the syntactic patterns relevant to measuring syntactic flexibility.[5]

Determiner type. A strong correlation has been observed between the flexibility of the determiner preceding the noun in a verb+noun combination and the overall flexibility of the phrase (Fellbaum 1993; Kearns 2002; Desbiens and Simon 2003). It is however

5 Note that there are idioms that appear primarily in a passivized form, for example, *the die is cast* ("the decision is made and will not change"). Our measure can in principle recognize such idioms because we do not require that an idiom appears mainly in active form; rather, we include voice (passive or active) as an important part of the syntactic pattern of an idiomatic combination.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

important to note that the nature of the determiner is also affected by other factors, such as the semantic properties of the noun. For this reason, determiner flexibility is sometimes argued not to be a good predictor of the overall syntactic flexibility of an expression. Nonetheless, many researchers consider it as an important part in the process of idiomatization of a verb+noun combination (Akimoto 1999; Kytö 1999; Tanabe 1999). We thus expect a VNIC to mainly appear with one type of determiner.

Pluralization. Although the verb constituent of a VNIC is morphologically flexible, the morphological flexibility of the noun relates to its referential status (Grant 2005). Again, one should note that the use of a singular or plural noun in a VNIC may also be affected by the semantic properties of the noun. Recall that during the idiomatization process, the noun constituent may become more abstract in meaning. In this process, the noun may lose some of its nominal features, including number (Akimoto 1999). The non-referential noun constituent of a VNIC is thus expected to mainly appear in just one of the singular or plural forms.

Merging the three types of variation results in a pattern set, $\mathcal{P}$, of 11 distinct syntactic patterns that are displayed in Table 1 along with examples for each pattern. When developing this set of patterns, we have taken into account the linguistic theories about the syntactic constraints on idiomatic expressions; for example, our choice of patterns is consistent with the idiom typology developed by Nicolas (1995). Note that we merge some of the individual patterns into one; for example, we include only one passive pattern independently of the choice of the determiner or the number of the noun. The motivation here is to merge low frequency patterns (i.e., those that are expected to be less common) in order to acquire more reliable evidence on the distribution of a particular verb–noun pair over the resulting pattern set. In principle, however, the set can be expanded to include more patterns; it can also be modified to contain different patterns for different classes of idiomatic combinations.

3.2.2 Devising a Statistical Measure. The second step is to devise a statistical measure that quantifies the degree of syntactic fixedness of a verb–noun pair, with respect to

Table 1
Patterns used in the syntactic fixedness measure, along with examples for each. A pattern signature is composed of a verb v in active (v_{act}) or passive (v_{pass}) voice; a determiner (det) that can be NULL, indefinite (*a/an*), definite (*the*), demonstrative (DEM), or possessive (POSS); and a noun n that can be singular (n_{sg}) or plural (n_{pl}).

Pattern No.	Pattern Signature			Example
1	v_{act}	det:NULL	n_{sg}	*give money*
2	v_{act}	det:*a/an*	n_{sg}	*give a book*
3	v_{act}	det:*the*	n_{sg}	*give the book*
4	v_{act}	det:DEM	n_{sg}	*give this book*
5	v_{act}	det:POSS	n_{sg}	*give my book*
6	v_{act}	det:NULL	n_{pl}	*give books*
7	v_{act}	det:*the*	n_{pl}	*give the books*
8	v_{act}	det:DEM	n_{pl}	*give those books*
9	v_{act}	det:POSS	n_{pl}	*give my books*
10	v_{act}	det:OTHER	$n_{sg,pl}$	*give many books*
11	v_{pass}	det:ANY	$n_{sg,pl}$	*a/the/this/my book/books was/were given*

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

the selected set of patterns, $\mathcal{P}$. We propose a measure that compares the syntactic behavior of the target pair with that of a "typical" verb–noun pair. Syntactic behavior of a typical pair is defined as the prior probability distribution over the patterns in $\mathcal{P}$. The maximum likelihood estimate for the prior probability of an individual pattern $pt \in \mathcal{P}$ is calculated as

$$P(pt) = \frac{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} f(v_i, n_j, pt)}{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} \sum_{pt_k \in \mathcal{P}} f(v_i, n_j, pt_k)} = \frac{f(*, *, pt)}{f(*, *, *)} \quad (3)$$

where $\mathcal{V}$ is the set of all instances of transitive verbs in the corpus, and $\mathcal{N}$ is the set of all instances of nouns appearing as the direct object of some verb.

The syntactic behavior of the target verb–noun pair $\langle v, n \rangle$ is defined as the posterior probability distribution over the patterns, given the particular pair. The maximum likelihood estimate for the posterior probability of an individual pattern pt is calculated as

$$P(pt \mid v, n) = \frac{f(v, n, pt)}{\sum_{pt_k \in \mathcal{P}} f(v, n, pt_k)} = \frac{f(v, n, pt)}{f(v, n, *)}. \quad (4)$$

The degree of syntactic fixedness of the target verb–noun pair is estimated as the divergence of its syntactic behavior (the posterior distribution over the patterns) from the typical syntactic behavior (the prior distribution). The divergence of the two probability distributions is calculated using a standard information-theoretic measure, the Kullback Leibler (KL-) divergence (Cover and Thomas 1991):

$$\text{Fixedness}_{\text{syn}}(v, n) \doteq D(P(pt \mid v, n) \,||\, P(pt)) = \sum_{pt_k \in \mathcal{P}} P(pt_k \mid v, n) \log \frac{P(pt_k \mid v, n)}{P(pt_k)} \quad (5)$$

KL-divergence has proven useful in many NLP applications (Resnik 1999; Dagan, Pereira, and Lee 1994). KL-divergence is always non-negative and is zero if and only if the two distributions are exactly the same. Thus, $\text{Fixedness}_{\text{syn}}(v, n) \in [0, +\infty]$, where large values indicate higher degrees of syntactic fixedness.

3.3 A Unified Measure of Fixedness

VNICs are hypothesized to be, in most cases, both lexically and syntactically more fixed than literal verb+noun combinations (see Section 2). We thus propose a new measure

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

of idiomaticity to be a measure of the overall fixedness of a given pair. We define $\text{Fixedness}_{\text{overall}}(v, n)$ as a weighted combination of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$:

$$\text{Fixedness}_{\text{overall}}(v, n) \doteq \alpha\, \text{Fixedness}_{\text{syn}}(v, n) + (1 - \alpha)\, \text{Fixedness}_{\text{lex}}(v, n) \quad (6)$$

where α weights the relative contribution of the measures in predicting idiomaticity.

Recall that $\text{Fixedness}_{\text{lex}}(v, n) \in [-\infty, +\infty]$, and $\text{Fixedness}_{\text{syn}}(v, n) \in [0, +\infty]$. To combine them in the overall fixedness measure, we rescale them, so that they fall in the range [0, 1]. Thus, $\text{Fixedness}_{\text{overall}}(v, n) \in [0, 1]$, where values closer to 1 indicate a higher degree of overall fixedness.

4. VNIC Type Recognition: Evaluation

To evaluate our proposed fixedness measures, we analyze their appropriateness for determining the degree of idiomaticity of a set of experimental expressions (in the form of verb–noun pairs, extracted as described in Section 4.1). More specifically, we first use each measure to assign scores to the experimental pairs. We then use the scores assigned by each measure to perform two different tasks, and assess the overall goodness of the measure by looking at its performance in both.

First, we look into the classification performance of each measure by using the scores to separate idiomatic verb–noun pairs from literal ones in a mixed list. This is done by setting a threshold, here the median score, where all pairs with scores higher than the threshold are labeled as idiomatic and the rest as literal.[6] For classification, we report accuracy (*Acc*), as well as the relative error rate reduction (*ERR*) over a random (chance) baseline, referred to as Rand. Second, we examine the retrieval performance of our fixedness measures by using the scores to rank verb–noun pairs according to their degree of idiomaticity. For retrieval, we present the precision–recall curves, as well as the interpolated three-point average precision or *IAP*—that is, the average of the interpolated precisions at the recall levels of 20%, 50%, and 80%. The interpolated average precision and precision–recall curves are commonly used for the evaluation of information retrieval systems (Manning and Schütze 1999), and reflect the goodness of a measure in placing the relevant items (here, idioms) before the irrelevant ones (here, literals).

Idioms are often assumed to exhibit collocational behavior to some extent, that is, the components of an idiom are expected to appear together more often than expected by chance. Hence, some NLP systems have used collocational measures to identify them (Smadja 1993; Evert and Krenn 2001). However, as discussed in Section 2, idioms have distinctive syntactic and semantic properties that separate them from simple collocations. For example, although collocations involve some degree of semantic idiosyncrasy (*strong tea* vs. ?*powerful tea*), compared to idioms, they typically have a more transparent meaning, and their syntactic behavior is more similar to that of literal expressions. We thus expect our fixedness measures that draw on the distinctive linguistic properties of idioms to be more appropriate than measures of collocation for the identification of idioms. To verify this hypothesis, in both the classification and retrieval tasks, we compare the performance of the fixedness measures with that of two collocation extraction measures: an informed baseline, PMI, and a position-based fixedness measure proposed

6 We adopt the median for this particular (balanced) data set, understanding that in practice a suitable threshold would need to be determined, e.g., based on development data.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

by Smadja (1993), which we refer to as Smadja. Next, we provide more details on PMI and Smadja.

PMI is a widely used measure for extracting statistically significant combinations of words or collocations. It has also been used for the recognition of idioms (Evert and Krenn 2001), warranting its use as an informed baseline here for comparison.[7] As in Equation (1), our calculation of PMI here restricts the counts of the verb–noun pair to the direct object relation. Smadja (1993) proposes a collocation extraction method which measures the fixedness of a word sequence (e.g., a verb–noun pair) by examining the relative position of the component words across their occurrences together. We replicate Smadja's method, where we measure fixedness of a target verb–noun pair as the spread (variance) of the co-occurrence frequency of the verb and the noun over 10 relative positions within a five-word window.[8]

Recall from Section 3.1 that our $\text{Fixedness}_{\text{lex}}$ measure is intended as an improvement over the non-compositionality measure of Lin (1999). For the sake of completeness, we also compare the classification performance of our $\text{Fixedness}_{\text{lex}}$ with that of Lin's (1999) measure, which we refer to as Lin.[9]

We first elaborate on the methodological aspects of our experiments in Section 4.1, and then present a discussion of the experimental results in Section 4.2.

4.1 Experimental Setup

4.1.1 Corpus and Data Extraction. We use the British National Corpus (BNC; Burnard 2000); to extract verb–noun pairs, along with information on the syntactic patterns they appear in. We automatically parse the BNC using the Collins parser (Collins 1999), and augment it with information about verb and noun lemmas, automatically generated using WordNet (Fellbaum 1998). We further process the corpus using TGrep2 (Rohde 2004) in order to extract syntactic dependencies. For each instance of a transitive verb, we use heuristics to extract the noun phrase (NP) in either the direct object position (if the sentence is active), or the subject position (if the sentence is passive). We then automatically find the head noun of the extracted NP, its number (singular or plural), and the determiner introducing it.

4.1.2 Experimental Expressions. We select our development and test expressions from verb–noun pairs that involve a member of a predefined list of transitive verbs, referred to as **basic verbs**. Basic verbs, in their literal use, refer to states or acts that are central to human experience. They are thus frequent, highly polysemous, and tend to combine with other words to form idiomatic combinations (Cacciari 1993; Claridge 2000; Gentner and France 2004). An initial list of such verbs was selected from several linguistic and psycholinguistic studies on basic vocabulary (Ogden 1968; Clark 1978; Nunberg, Sag, and Wasow 1994; Goldberg 1995; Pauwels 2000; Claridge 2000; Newman and Rice 2004). We further augmented this initial list with verbs that are semantically related to another

7 PMI has been shown to perform better than or comparable to many other association measures (Inkpen 2003; Mohammad and Hirst, submitted). In our experiments, we also found that PMI consistently performs better than two other association measures, the Dice coefficient and the log-likelihood measure. Experiments by Krenn and Evert (2001) showed contradicting results for PMI; however, these experiments were performed on small-sized corpora, and on data which contained items with very low frequency.

8 We implement the method as explained in Smadja (1993), taking into account the part-of-speech tags of the target component words.

9 We implement the method as explained in Lin (1999), using 95% confidence intervals. We thus need to ignore variants with frequency lower than 4 for which no confidence interval can be formed.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

verb already in the list; for example, *lose* is added in analogy with *find*. Here is the final list of the 28 verbs in alphabetical order:

blow, bring, catch, cut, find, get, give, have, hear, hit, hold, keep, kick, lay, lose, make, move, place, pull, push, put, see, set, shoot, smell, take, throw, touch

From the corpus, we extract all the verb–noun pairs (lemmas) that contain any of these listed basic verbs, and that appear at least 10 times in the corpus in a direct object relation (irrespective of any intervening determiners or adjectives). From these, we select a subset that are idiomatic, and another subset that are literal, as follows: A verb–noun pair is considered idiomatic if it appears in an idiom listed in a credible dictionary such as the *Oxford Dictionary of Current Idiomatic English* (ODCIE; Cowie, Mackin, and McCaig 1983), or the Collins COBUILD Idioms Dictionary (CCID; Seaton and Macaulay 2002).[10] To decide whether a verb–noun pair has appeared in an idiom, we look for all idioms containing the verb and the noun in a direct-object relation, irrespective of any intervening determiners or adjectives, and/or any other arguments. The pair is considered literal if it involves a physical act or state (i.e., the basic semantics of the verb) and does not appear in any of the mentioned dictionaries as an idiom (or part of an idiom). From the set of idiomatic pairs, we then randomly pull out 80 development pairs and 100 test pairs, ensuring that we have items of both low and high frequency. We then double the size of each data set (development and test) by adding equal numbers of literal pairs, with similar frequency distributions. Some of the idioms corresponding to the experimental idiomatic pairs are: *kick the habit*, *move mountains*, *lose face*, and *keep one's word*. Examples of literal pairs include: *move carriage*, *lose ticket*, and *keep fish*.

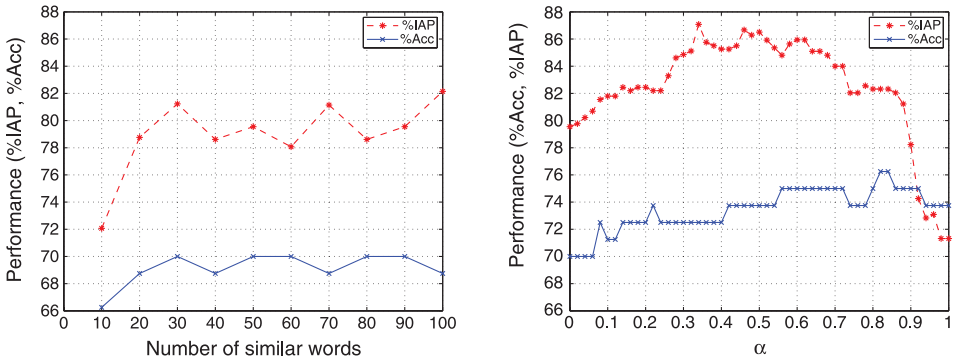
Development expressions are used in devising the fixedness measures, as well as in determining the values of their parameters as explained in the next subsection. Test expressions are saved as unseen data for the final evaluation.

4.1.3 Parameter Settings. Our lexical fixedness measure in Equation (2) involves two parameters, K_v and K_n, which determine the number of lexical variants considered in measuring the lexical fixedness of a given verb–noun pair. We make the least-biased assumption on the proportion of variants generated by replacing the verb (K_v) and those generated by replacing the noun (K_n)—that is, we assume $K_v = K_n$.[11] We perform experiments on the development data, where we set the total number of variants (i.e., $K_v + K_n$) from 10 to 100 by steps of 10. (For simplicity, we refer to the total number of variants as K). Figure 1(a) shows the change in performance of $\text{Fixedness}_{\text{lex}}$ as a function of K. Recall that *Acc* is the classification accuracy, and *IAP* reflects the average precision of a measure in ranking idiomatic pairs before non-idiomatic ones. According to these results, there is not much variation in the performance of the measure for

10 Our development data also contains items from several other dictionaries, such as *Chambers Idioms* (Kirkpatrick and Schwarz 1982). However, our test data, which is also used in the token-based experiments, however, only contains idioms from the two dictionaries ODCIE and CCID. Results reported in this article are all on test pairs; development pairs are mainly used for the development of the methods.

11 We also performed experiments on the development data in which we did not restrict the number of variants, and hence did not enforce the condition $K_v = K_n$. Instead, we tried using a variety of thresholds on the similarity scores (from the thesaurus) in order to find the set of most similar words to a given verb or noun. We found that fixing the number of most similar words is more effective than using a similarity threshold, perhaps because the actual scores can be very different for different words.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024



(a) Performance of $\text{Fixedness}_{\text{lex}}$ as a function of K.

(b) Performance of $\text{Fixedness}_{\text{overall}}$ as a function of α.

Figure 1
%*IAP* and %*Acc* of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{overall}}$ over development data.

$K \geq 20$. We thus choose an intermediate value for K that yields the highest accuracy and a reasonably high precision; specifically, we set K to 50.

The overall fixedness measure defined in Equation (6) also uses a parameter, α, which determines the relative weights given to the individual fixedness measures in the linear combination. We experiment on the development data with different values of α ranging from 0 to 1 by steps of .02; results are shown in Figure 1(b). As can be seen in the figure, the accuracy of $\text{Fixedness}_{\text{overall}}$ is not affected much by the change in the value of α. The average precision (*IAP*), however, shows that the combined measure performs best when somewhat equal weights are given to the two individual measures, and performs worst when the lexical fixedness component is completely ignored (i.e., α is close to 1). These results also reinforce that a complete evaluation of our fixedness measures should include both metrics, accuracy, and average precision, as they reveal different aspects of performance. Here, for example, $\text{Fixedness}_{\text{syn}}$ ($\alpha = 1$) has comparable accuracy to $\text{Fixedness}_{\text{lex}}$ ($\alpha = 0$), reflecting that the two measures generally give higher scores to idioms. However, the ranking precision of the latter is much higher than that of the former, showing that $\text{Fixedness}_{\text{lex}}$ ranks many of the idioms at the very top of the list. In all our experiments reported here, we set α to .6, a value for which $\text{Fixedness}_{\text{overall}}$ shows reasonably good performance according to both *Acc* and *IAP*.

4.2 Experimental Results and Analysis

In this section, we report the results of evaluating our measures on unseen test expressions, with parameters set to the values determined in Section 4.1.3. (Results on development data have similar trends to those on test data.) We analyze the classification performance of the individual lexical and syntactic fixedness measures in Section 4.2.1, and discuss their effectiveness for retrieval in Section 4.2.2. Section 4.2.3 then looks into the performance of the overall fixedness measure, and Section 4.2.4 presents a summary and discussion of the results.

4.2.1 Classification Performance. Here, we look into the performance of the individual fixedness measures, $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$, in classifying a mixed set of verb–noun pairs into idiomatic and literal classes. We compare their performance against the

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 2
Accuracy and relative error reduction for the two fixedness measures, the two baseline measures, and Smadja, over all test pairs (TEST_{all}), and test pairs divided by frequency ($\text{TEST}_{f_{\text{low}}}$ and $\text{TEST}_{f_{\text{high}}}$).

	TEST_{all}		$\text{TEST}_{f_{\text{low}}}$		$\text{TEST}_{f_{\text{high}}}$	
Measure	%*Acc*	(%*ERR*)	%*Acc*	(%*ERR*)	%*Acc*	(%*ERR*)
Rand	50		50		50	
PMI	63	(26)	56	(12)	70	(40)
Smadja	54	(8)	64	(28)	62	(24)
$\text{Fixedness}_{\text{lex}}$	68	(36)	70	(40)	70	(40)
$\text{Fixedness}_{\text{syn}}$	**71**	(**42**)	**72**	(**44**)	**82**	(**64**)

two baselines, Rand and PMI, as well as the two state-of-the-art methods, Smadja and Lin. For analytical purposes, we further divide the set of all test expressions, TEST_{all}, into two sets corresponding to two frequency bands: $\text{TEST}_{f_{\text{low}}}$ contains 50 idiomatic and 50 literal pairs, each with total frequency (across all syntactic patterns under consideration) between 10 and 40; $\text{TEST}_{f_{\text{high}}}$ consists of 50 idiomatic and 50 literal pairs, each with total frequency of 40 or greater. Classification performances of all measures except Lin are given in Table 2. Lin does not assign scores to the test verb–noun pairs, hence we cannot calculate its classification accuracy the same way we do for the other methods (i.e., using median as the threshold). A separate comparison between Lin and $\text{Fixedness}_{\text{lex}}$ is provided at the end of this section.

As can be seen in the first two columns of Table 2, the informed baseline, PMI, shows a large improvement over the random baseline (26% error reduction) on TEST_{all}. This shows that many VNICs have turned into institutionalized (i.e., statistically significant) co-occurrences. Hence, one can get relatively good performance by treating verb+noun idiomatic combinations as collocations. $\text{Fixedness}_{\text{lex}}$ performs considerably better than the informed baseline (36% vs. 26% error reduction on TEST_{all}). $\text{Fixedness}_{\text{syn}}$ has the best performance (shown in boldface), with 42% error reduction over the random baseline, and 21.6% error reduction over PMI. These results demonstrate that lexical and syntactic fixedness are good indicators of idiomaticity, better than a simple measure of collocation such as PMI. On TEST_{all}, Smadja performs only slightly better than the random baseline (8% error reduction), reflecting that a position-based fixedness measure is not sufficient for identifying idiomatic combinations. These results suggest that looking into deep linguistic properties of VNICs is necessary for the appropriate treatment of these expressions.[12]

PMI is known to perform poorly on low frequency items. To examine the effect of frequency on the measures, we analyze their performance on the two divisions of the

12 Performing the χ^2 test of statistical significance, we find that the differences between Smadja and our lexical and syntactic fixedness measures are statistically significant at $p < 0.05$. However, the differences in performance between fixedness measures and PMI are not statistically significant. Note that this does not imply that the differences are not substantial, rather that there is not enough evidence in the observed data to reject the null hypothesis (that two methods perform the same in general) with high confidence. Moreover, χ^2 is a non-parametric (distribution free) test and hence it has less power to reject a null hypothesis. Later, when we take into account the actual scores assigned by the measures, we find that all differences are statistically significant (see Sections 4.2.2–4.2.3 for more details). All significance tests are performed using the *R* (2004) package.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

test data, corresponding to the two frequency bands, $\text{TEST}_{f_{\text{low}}}$ and $\text{TEST}_{f_{\text{high}}}$. Results are given in the four rightmost columns of Table 2, with the best performance shown in boldface. As expected, the performance of PMI drops substantially for low frequency items. Interestingly, although it is a PMI-based measure, $\text{Fixedness}_{\text{lex}}$ has comparable performance on all data sets. The performance of $\text{Fixedness}_{\text{syn}}$ improves quite a bit when it is applied to high frequency items, while maintaining similar performance on the low frequency items. These results show that the lexical and syntactic fixedness measures perform reasonably well on both low and high frequency items.[13] Hence they can be used with a higher degree of confidence, especially when applied to data that is heterogeneous with regard to frequency. This is important because, while some VNICs are very common, others have very low frequency, as noted by Grant (2005). Smadja shows a notable improvement in performance when data is divided by frequency. This effect is likely due to the fact that fixedness is measured as the spread of the position-based (raw) co-occurrence frequencies. Nonetheless, on both data sets the performance of Smadja remains substantially worse than that of our two fixedness measures (the differences are statistically significant in three out of the four comparisons at $p < .05$).

Collectively, these results show that our linguistically motivated fixedness measures are particularly suited for identifying idiomatic combinations, especially in comparison with more general collocation extraction techniques, such as PMI or the position-based fixedness measure of Smadja (1993). Especially, our measures tend to perform well on low frequency items, perhaps due to their reliance on distinctive linguistic properties of idioms.

We now compare the classification performance of $\text{Fixedness}_{\text{lex}}$ to that of Lin. Unlike $\text{Fixedness}_{\text{lex}}$, Lin does not assign continuous scores to the verb–noun pairs, but rather classifies them as idiomatic or non-idiomatic. Thus, we cannot use the same threshold (e.g., median) for the two methods to calculate their classification accuracies in a comparable way. Recall also from Section 3.1 that the performance of both these methods depends on the value of K (the number of variants). We thus measure the classification precision of the methods at equivalent levels of recall, using the same number of variants K at each recall level for the two measures. Varying K from 2 to 100 by steps of 4, Lin and $\text{Fixedness}_{\text{lex}}$ achieve an average classification precision of 81.5% and 85.8%, respectively. Performing a t-test on the precisions of the two methods confirms that the difference between the two is statistically significant at $p < .001$. In addition, our method has the advantage of assigning a score to a target verb–noun reflecting its *degree* of lexical fixedness. Such information can help a lexicographer decide whether a given verb–noun should be placed in a lexicon.

4.2.2 Retrieval Performance. The classification results suggest that the individual fixedness measures are overall better than a simple measure of collocation at separating idiomatic pairs from literal ones. Here, we have a closer look at their performance by examining their goodness in ranking verb–noun pairs according to their degree of idiomaticity. Recall that the fixedness measures are devised to reflect the degree of fixedness and hence the degree of idiomaticity of a target verb–noun pair. Thus, the result of applying each measure to a list of mixed pairs is a list that is ranked in the order

13 In fact, the results show that the performance of both fixedness measures is better when data is divided by frequency. Although we expect better performance over high frequency items, more investigation is needed to verify whether the improvement in performance over low frequency items is a meaningful effect or merely an accident of the data at hand.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

of idiomaticity. For a measure to be considered good at retrieval, we expect idiomatic pairs to be very frequent near the top of the ranked list, and to become less frequent towards the bottom. Precision–recall curves are very indicative of this trend: The ideal measure will have a precision of 100% for all values of recall, namely, the measure places all idiomatic pairs at the very top of the ranked list. In reality, although the precision drops as recall increases, we expect a good measure to keep high precision at most levels of recall.

Figure 2 depicts the interpolated precision–recall curves for PMI and Smadja, and for the lexical, syntactic, and overall fixedness measures, over TEST_{all}. Note that the minimum interpolated precision is 50% due to the equal number of idiomatic and literal pairs in the test data. In this section, we discuss the retrieval performance of the two individual fixedness measures; the next section analyzes the performance of the overall fixedness measure.

The precision–recall curves of Smadja and PMI are nearly flat (with PMI consistently higher than Smadja), showing that the distribution of idiomatic pairs in the lists ranked by these two measures is only slightly better than random. A close look at the precision–recall curve of $\text{Fixedness}_{\text{lex}}$ reveals that, up to the recall level of 50%, the precision of this measure is substantially higher than that of PMI. This means that, compared to PMI, $\text{Fixedness}_{\text{lex}}$ places more idiomatic pairs at the very top of the list. At higher recall levels (50% and higher), $\text{Fixedness}_{\text{lex}}$ still consistently outperforms PMI. Nonetheless, at these recall values, the two measures have relatively low precision (compared to the other measures), suggesting that both measures also put many idiomatic pairs near the bottom of the list. In contrast, the precision–recall curve of $\text{Fixedness}_{\text{syn}}$ shows that its performance is consistently much better than that of PMI: Even at the recall level of 90%, its precision is close to 70% (cf. 55% precision of PMI).

A comparison of the precision–recall curves of the two individual fixedness measures reveals their complementary nature. Compared to $\text{Fixedness}_{\text{lex}}$, $\text{Fixedness}_{\text{syn}}$ maintains higher precision at very high levels of recall, suggesting that the syntactic fixedness measure places fewer idiomatic pairs at the bottom of the ranked list. In contrast, $\text{Fixedness}_{\text{lex}}$ has notably higher precision than $\text{Fixedness}_{\text{syn}}$ at recall levels of up to 40%, suggesting that the former puts more idiomatic pairs at the top of the ranked list.

Statistical significance tests confirm these observations: Using the Wilcoxon Signed Rank test (1945), we find that both $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$ produce significantly different rankings from PMI and Smadja ($p \ll .001$). Also, the rankings of the items

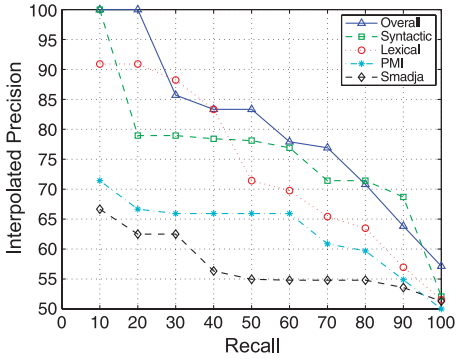


Figure 2
Precision–recall curves for PMI, Smadja, and for the fixedness measures, over TEST_{all}.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 3
Classification and retrieval performance of the overall fixedness measure over TEST_{all}.

Measure	%*Acc*	(%*ERR*)	%*IAP*
PMI	63	(26)	63.5
Smadja	54	(8)	57.2
$\text{Fixedness}_{\text{lex}}$	68	(36)	75.3
$\text{Fixedness}_{\text{syn}}$	71	(42)	75.9
$\text{Fixedness}_{\text{overall}}$	**74**	(**48**)	**84.7**

produced by the two individual fixedness measures are found to be significantly different at $p < .01$.

4.2.3 Performance of the Overall Fixedness Measure. We now look at the classification and retrieval performance of the overall fixedness measure. Table 3 presents %*Acc*, %*ERR*, and %*IAP* of $\text{Fixedness}_{\text{overall}}$, repeating that of PMI, Smadja, $\text{Fixedness}_{\text{lex}}$, and $\text{Fixedness}_{\text{syn}}$, for comparison. Here again the error reductions are relative to the random baseline of 50%. Looking at classification performance (expressed in terms of %*Acc* and %*ERR*), we can see that $\text{Fixedness}_{\text{overall}}$ notably outperforms all other measures, including lexical and syntactic fixedness (18.8% error reduction relative to $\text{Fixedness}_{\text{lex}}$, and 10% error reduction relative to $\text{Fixedness}_{\text{syn}}$). According to the classification results, each of the lexical and syntactic fixedness measures are good at separating idiomatic from literal combinations, with syntactic fixedness performing better. Here we demonstrate that combining them into a single measure of fixedness, while giving more weight to the better measure, results in a more effective classifier.[14] The overall behavior of this measure as a function of α is displayed in Figure 3.

As can be seen in Table 3, $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$ have comparable *IAP*: 75.3% and 75.9%, respectively. In comparison, $\text{Fixedness}_{\text{overall}}$ has a much higher *IAP* of 84.7%, reinforcing the claim that combining evidence from both lexical and syntactic fixedness is beneficial. Recall from Section 4.2.2 that the two individual fixedness measures exhibit complementary behavior, as observed in their precision–recall curves shown in Figure 2. The precision–recall curve of the overall fixedness measure shows that this measure in fact combines advantages of the two individual measures: At most recall levels, $\text{Fixedness}_{\text{overall}}$ has a higher precision than both individual measures. Statistical significance tests that look at the actual scores assigned by the measures confirm that the observed differences in performance are significant. The Wilcoxon Signed Rank test shows that the $\text{Fixedness}_{\text{overall}}$ measure produces a ranking that is significantly different from those of the individual fixedness measures, the baseline PMI, and Smadja (at $p \ll .001$).

4.2.4 Summary and Discussion. Overall, the worst performance belongs to the two collocation extraction methods, PMI and Smadja, both in classifying test pairs as idiomatic or

14 Using a χ^2 test, we find a statistically significant difference between the classification performance of $\text{Fixedness}_{\text{overall}}$ and that of Smadja ($p < 0.01$), and also a marginally significant difference between the performance of $\text{Fixedness}_{\text{overall}}$ and that of PMI ($p < .1$). Recall from footnote 12 (page 15) that none of the individual measures' performances significantly differed from that of PMI. Nonetheless, no significant differences are found between the classification performance of $\text{Fixedness}_{\text{overall}}$ and that of the individual fixedness measures.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

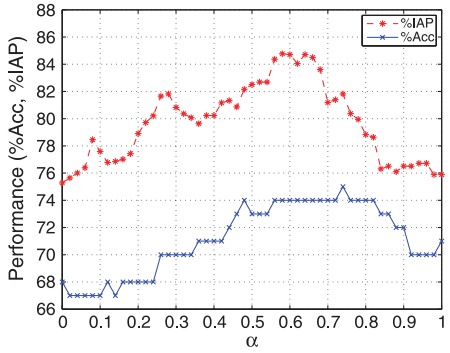


Figure 3
Classification performance of $\text{Fixedness}_{\text{overall}}$ on test data as a function of α.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

literal, and in ranking the pairs according to their degree of idiomaticity. This suggests that although some VNICs are institutionalized, many do not appear with markedly high frequency, and hence only looking at their frequency is not sufficient for their recognition. Moreover, a position-based fixedness measure does not seem to sufficiently capture the syntactic fixedness of VNICs in contrast to the flexibility of literal phrases. $\text{Fixedness}_{\text{overall}}$ is the best performer of all, supporting the hypothesis that many VNICs are both lexically and syntactically fixed, more so than literal verb+noun combinations. In addition, these results demonstrate that incorporating such linguistic properties into statistical measures is beneficial for the recognition of VNICs.

Although we focus on experimental expressions with frequency higher than 10, PMI still shows great sensitivity to frequency differences, performing especially poorly on items with frequency between 10 and 40. In contrast, none of the fixedness measures are as sensitive to such frequency differences. Especially interesting is the consistent performance of $\text{Fixedness}_{\text{lex}}$, which is a PMI-based measure, on low and high frequency items. These observations put further emphasis on the importance of devising new methods for extracting multiword expressions with particular syntactic and semantic properties, such as VNICs.

To further analyze the performance of the fixedness measures, we look at the top and bottom 20 pairs (10%) in the lists ranked by each fixedness measure. Interestingly, the list ranked by $\text{Fixedness}_{\text{overall}}$ contains no false positives (*fp*) in the top 20 items, and no false negatives (*fn*) in the bottom 20 items, once again reinforcing the usefulness of combining evidence from the individual lexical and syntactic fixedness measures. False positive and false negative errors found in the top and bottom 20 ranked pairs, respectively, for the syntactic and lexical fixedness measures are given in Table 4. (Note that *fp* errors are the non-idiomatic pairs ranked at the top, whereas *fn* errors are the idiomatic pairs ranked at the bottom.)

We first look at the errors made by $\text{Fixedness}_{\text{syn}}$. The first *fp* error, *throw hat*, is an interesting one: even though the pair is not an idiomatic expression on its own, it is part of the larger idiomatic phrase *throw one's hat in the ring*, and hence exhibits syntactic fixedness. This shows that our methods can be easily extended to identify other types of verb phrase idiomatic combinations which exhibit syntactic behavior similar to VNICs. Looking at the frequency distribution of the occurrence of the other two *fp* errors, *touch finger* and *lose home*, in the 11 patterns from Table 1, we observe that both pairs tend to appear mainly in the patterns "v_{act} det:POSS n_{sg}" (*touch one's finger*, *lose one's home*) and/or "v_{act} det:POSS n_{pl}" (*touch one's fingers*). These examples show

Table 4
Errors found in the top and bottom 20 pairs in the lists ranked by the two individual fixedness measures; *fp* stands for false positive, *fn* stands for false negative.

Measure:	$\text{Fixedness}_{\text{syn}}$		$\text{Fixedness}_{\text{lex}}$	
Error Type:	*fp*	*fn*	*fp*	*fn*
	throw hat *touch finger* *lose home*	*make pile* *keep secret*	*push barrow* *blow bridge*	*have moment* *give way* *keep hand*

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

that syntactic fixedness is not a sufficient condition for idiomaticity. In other words, it is possible for non-idiomatic expressions to be syntactically fixed for reasons other than semantic idiosyncrasy. In these examples, the nouns *finger* and *home* tend to be introduced by a possessive determiner, because they often belong to someone. It is also important to note that these two patterns have a low prior (i.e., verb–noun pairs do not typically appear in these patterns). Hence, an expression with a strong tendency to appear in such patterns will be given a high syntactic fixedness score.

The frequency distribution of the two *fn* errors for $\text{Fixedness}_{\text{syn}}$ reveals that they are given low scores mainly because their distributions are similar to the prior. Even though *make pile* preferably appears in the two patterns "v_{act} det:*a/an* n_{sg}" and "v_{act} det:NULL n_{pl}," both patterns have reasonably high prior probabilities. Moreover, because of the low frequency of *make pile* (< 40), the evidence is not sufficient to distinguish it from a typical verb–noun pair. The pair *keep secret* has a high frequency, but its occurrences are scattered across all 11 patterns, closely matching the prior distribution. The latter example shows that syntactic fixedness is not a necessary condition for idiomaticity either.[15]

Analyzing the errors made by $\text{Fixedness}_{\text{lex}}$ is more difficult as many factors may affect scores given by this measure. Most important is the quality of the automatically generated variants. We find that in one case, *push barrow*, the first 25 distributionally similar nouns (taken from the automatically built thesaurus) are proper nouns, perhaps because *Barrow* is a common last name. In general, it seems that the similar verbs and nouns for a target verb–noun pair are not necessarily related to the same sense of the target word. Another possible source of error is that in this measure we use PMI as an indirect clue to idiomaticity. In the case of *give way* and *keep hand*, many of the variants are plausible combinations with very high frequency of occurrence, for example, *give opportunity*, *give order*, *find way* for the former, and *hold hand*, *put hand*, *keep eye* for the latter. Whereas some of these high-frequency variants are literal (e.g., *hold hand*) or idiomatic (e.g., *keep eye*), many have metaphorical interpretations (e.g., *give opportunity*, *find way*). In our ongoing work, we use lexical and syntactic fixedness measures, in combination with other linguistically motivated features, to distinguish such metaphorical combinations from both literal and idiomatic expressions (Fazly and Stevenson, to appear).

One way to decrease the likelihood of making any of these errors is to combine evidence from the lexical and syntactic fixedness of idioms. As can be seen in Table 4, the two fixedness measures make different errors, and combining them results in a measure

15 One might argue that *keep secret* is more semantically analyzable and hence less idiomatic than an expression such as *shoot the breeze*. Nonetheless, it is still semantically more idiosyncratic than a fully literal combination such as *keep a pen*, and hence should not be ranked at the very bottom of the list.

(the overall fixedness) that makes fewer errors. In the future, we intend to also look into other properties of idioms, such as their semantic non-compositionality, as extra sources of information.

5. Determining the Canonical Forms of VNICs

Our evaluation of the fixedness measures demonstrates their usefulness for the automatic recognition of VNICs. Recall from Section 2 that idioms appear in restricted syntactic forms, often referred to as their canonical forms (Glucksberg 1993; Riehemann 2001; Grant 2005). For example, the idiom *pull one's weight* mainly appears in this form (when used idiomatically). The lexical representation of an idiomatic combination thus must contain information about its canonical forms. Such information is necessary both for automatically generating appropriate forms (e.g., in a natural language generation system or a machine translation system), and for inclusion in dictionaries for learners (e.g., in the context of computational lexicography).

Because VNICs are syntactically fixed, they are mostly expected to have a small number of canonical forms. For example, *shoot the breeze* is listed in many idiom dictionaries as the canonical form for $\langle shoot, breeze \rangle$. Also, *hold fire* and *hold one's fire* are listed in CCID as canonical forms for $\langle hold, fire \rangle$. We expect a VNIC to occur in its canonical form(s) with substantially higher frequency than in any other syntactic patterns. We thus devise an unsupervised method that discovers the canonical form(s) of a given idiomatic verb–noun pair by examining its frequency of occurrence in each syntactic pattern under consideration. Specifically, the set of the canonical form(s) of the target pair $\langle v, n \rangle$ is defined as

$$\mathcal{C}(v, n) = \{pt_k \in \mathcal{P} \mid z(v, n, pt_k) > T_z\} \tag{7}$$

Here, $\mathcal{P}$ is the set of patterns (see Table 1), and the condition $z(v, n, pt_k) > T_z$ determines whether the frequency of the target pair $\langle v, n \rangle$ in pt_k is substantially higher than its frequency in other patterns; $z(v, n, pt_k)$ is calculated using the statistic z-score as in Equation (8), and T_z is a predefined threshold.

$$z(v, n, pt_k) = \frac{f(v, n, pt_k) - \bar{f}}{s} \tag{8}$$

where $\bar{f}$ is the sample mean and s the sample standard deviation.

The statistic $z(v, n, pt_k)$ indicates how far and in which direction the frequency of occurrence of the target pair $\langle v, n \rangle$ in a particular pattern pt_k deviates from the sample mean, expressed in units of the sample standard deviation. To decide whether pt_k is a canonical pattern for the target pair, we check whether its z-score, $z(v, n, pt_k)$, is greater than a threshold T_z. Here, we set T_z to 1, based on the distribution of z and through examining the development data.

We evaluate our unsupervised canonical form identification method by verifying its predicted forms against ODCIE and CCID. Specifically, for each of the 100 idiomatic pairs in TEST_{all}, we calculate the precision and recall of its predicted canonical forms (those whose z-scores are above T_z), compared to the canonical forms listed in the two dictionaries. The average precision across the 100 test pairs is 81.2%, and the average recall is 88% (with 68 of the pairs having 100% precision and 100% recall). Moreover, we

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

find that for the overwhelming majority of the pairs, 86%, the predicted canonical form with the highest z-score appears in the dictionary entry of the pair.

According to the entries in ODCIE and CCID, 93 out of the 100 idiomatic pairs in TEST_{all} have one canonical form. Our canonical form extraction method on average finds 1.2 canonical forms for these 100 pairs (one canonical form for 79 of them, two for 18, and three for 3 of these). Generally, our method tends to extract more canonical forms than listed in the dictionaries. This is a desired property, because idiom dictionaries often do not exhaustively list all canonical forms, but the most dominant ones. Examples of such cases include: *see the sights* for which our method also finds *see sights* as a canonical form, and *catch one's attention* for which our method also finds *catch the attention*. There are also cases where our method finds canonical forms for a given expression due to noise resulting from the use of the expression in a non-idiomatic sense. For example, for *hold one's horses*, our method also finds *hold the horse* and *hold the horses* as canonical forms. Similarly, for *get the bird*, our method also finds *get a bird*.

In a few cases (4 out of 100), our method finds fewer canonical forms than listed in the dictionaries. These are *catch the/one's imagination*, *have a/one's fling*, *make a/one's mark*, and *have a/the nerve*. For the first two of these, the z-score of the missed pattern is only slightly lower than our predefined threshold. In other cases (8 out of 100), none of the canonical forms extracted by our method match those in a dictionary. Some of these expressions also have a non-idiomatic sense which might be more dominant than the idiomatic usage. For example, for *give the push* and *give the flick*, our method finds *give a push* and *give a flick*, respectively, perhaps due to the common use of the latter forms as light verb constructions. For *make one's peace*, our method finds a different form, *make peace*, which seems a plausible canonical form; and moreover, the canonical form listed in the dictionaries (*make one's peace*) has a z-score which is only slightly lower than our threshold. There is also one case where our method finds a canonical form that corresponds to a different idiom using the same verb+noun: we find *lose touch* as a canonical form, whereas the dictionaries list an idiom with a different canonical form (*lose one's touch*) as the idiom with *lose* and *touch*.

In general, canonical forms extracted by our method are reasonably accurate, but may need to be further analyzed by a lexicographer to filter out incorrectly found patterns. Moreover, our method extracts new canonical forms for some expressions, which could be used to augment dictionaries.

6. Automatic Identification of VNIC Tokens

In previous sections, we have provided an analysis of the lexical and syntactic behavior of idiomatic expressions. We have shown that our proposed techniques that draw on such properties can successfully distinguish an idiomatic verb+noun combination (a VNIC type) such as *get the sack* from a non-idiomatic (literal) one such as *get the bag*. It is important, however, to note that a potentially idiomatic expression such as *get the sack* can also have a literal interpretation in a given context, as in *Joe got the sack from the top shelf*. This is true of many potential idioms, although the relative proportion of literal usages may differ from one expression to another. For example, an expression such as *see stars* is much more likely to have a literal interpretation than *get the sack* (according to our findings in the BNC). Identification of idiomatic tokens in context is thus necessary for a full understanding of text, and this will be the focus of Sections 6 and 7.

Recent studies addressing token identification for idiomatic expressions mainly perform the task as one of word sense disambiguation, and draw on the local context of

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

a token to disambiguate it. Such techniques either do not use any information regarding the linguistic properties of idioms (Birke and Sarkar 2006), or mainly focus on the property of non-compositionality (Katz and Giesbrecht 2006). Studies that do make use of deep linguistic information often handcode the knowledge into the systems (Uchiyama, Baldwin, and Ishizaki 2005; Hashimoto, Sato, and Utsuro 2006). Our goal is to develop techniques that draw on the specific linguistic properties of idioms for their identification, without the need for handcoded knowledge or manually labelled training data. Such unsupervised techniques can also help provide automatically labelled (noisy) training data to bootstrap (semi-)supervised methods.

In Sections 3 and 4, we showed that the lexical and syntactic fixedness of idioms is especially relevant to their type-based recognition. We expect such properties to also be relevant for their token identification. Moreover, we have shown that it is possible to learn about the fixedness of idioms in an unsupervised manner. Here, we propose unsupervised techniques that draw on the syntactic fixedness of idioms to classify individual tokens of a potentially idiomatic phrase as literal or idiomatic. We also put forward a classification technique that combines such information (in the form of noisy training data) with evidence from the local context of usages of an expression. In Section 6.1, we elaborate on the underlying assumptions of our token identification techniques. Section 6.2 then describes our proposed methods that draw on these assumptions to perform the task.

6.1 Underlying Assumptions

Although there may be fine-grained differences in meaning across the idiomatic usages of an expression, as well as across its literal usages, we assume that the idiomatic and literal usages correspond to two coarse-grained senses of the expression. We will refer then to each of the literal and idiomatic designations as a (coarse-grained) meaning of the expression, while acknowledging that each may have multiple fine-grained senses.

Recall from Section 2 that idioms tend to be somewhat fixed with respect to the syntactic configurations in which they occur. For example, *pull one's weight* tends to mainly appear in this form when used idiomatically. Other forms of the expression, such as *pull the weights*, typically are only used with a literal meaning. In other words, an idiom tends to have one (or a small number of) canonical form(s), which are its most preferred syntactic patterns.[16] Here we assume that, in most cases, idiomatic usages of an expression tend to occur in its canonical form(s). We also assume that, in contrast, the literal usages of an expression are less syntactically restricted, and are expressed in a greater variety of patterns. Because of their relative unrestrictedness, literal usages may occur in a canonical form for that expression, but usages in a canonical form are more likely to be idiomatic. Usages in alternative syntactic patterns for the expression, which we refer to as the non-canonical forms of the expression, are more likely to be literal.

Drawing on these assumptions, we develop unsupervised methods that determine, for each verb+noun token in context, whether it has an idiomatic or a literal

16 As noted previously, 93 out of the 100 idiomatic pairs in TEST_{all} have one canonical form, according to the entries in ODCIE and CCID. Also, our canonical form extraction method on average finds 1.2 canonical forms for the 100 test idioms.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

interpretation. Clearly, the success of our methods depends on the extent to which these assumptions hold (we will return to these assumptions in Section 7.2.3).

6.2 Proposed Methods

This section elaborates on our proposed methods for identifying the idiomatic and literal usages of a verb+noun combination: the CFORM method that uses knowledge of canonical forms only, and the CONTEXT method that also incorporates distributional evidence about the local context of a token. Both methods draw on our assumptions described herein, that usages in the canonical form(s) for a potential idiom are more likely to be idiomatic, and those in other forms are more likely to be literal. Because our methods need information about canonical forms of an expression, we use the unsupervised method described in Section 5 to find these automatically. In the following discussion, we describe each method in more detail.

CFORM. This method classifies an instance (token) of an expression as idiomatic if it occurs in one of the automatically determined canonical form(s) for that expression (e.g., *pull one's weight*), and as literal otherwise (e.g., *pull a weight*, *pull the weights*). The underlying assumption of this method is that information about the canonical form(s) of an idiom type can provide a reasonably accurate classification of its individual instances as literal or idiomatic.

CONTEXT. Recall our assumption that the idiomatic and literal usages of an idiom correspond to two coarse-grained meanings of the expression. It is natural to further assume that the literal and idiomatic usages have more in common semantically within each group than between the two groups. Adopting a distributional approach to meaning—where the meaning of an expression is approximated by the words with which it co-occurs (Firth 1957)—we would expect the literal and idiomatic usages of an expression to typically occur with different sets of words.

Indeed, in a supervised setting, Katz and Giesbrecht (2006) show that the local context of an idiom usage is useful in identifying its sense. Inspired by this work, we propose an unsupervised method that incorporates distributional information about the local context of the usages of an idiom, in addition to the (syntactic) knowledge about its canonical forms, in order to determine if its token usages are literal or idiomatic. To achieve this, the method compares the context surrounding a test instance of an expression to "gold-standard" contexts for the idiomatic and literal usages of the expression, which are taken from noisy training data automatically labelled using canonical forms.[17]

For each test instance of an expression, the CONTEXT method thus compares its co-occurring words to two sets of gold-standard co-occurring words: one typical of idiomatic usages and one typical of literal usages of the expression (we will shortly explain precisely how we find these). If the test token is determined to be (on average) more similar to the idiomatic usages, then it is labelled as idiomatic. Otherwise, it is labelled as literal. To measure similarity between two sets of words, we use

17 The two CONTEXT methods in our earlier work (Cook, Fazly, and Stevenson 2007) were biased because they used information about the canonical form of a test token (in addition to context information). We found that when the bias was removed, the similarity measure used in those techniques was not as effective, and hence we have developed a different method here.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

a standard distributional similarity measure, Jaccard, defined subsequently.[18] In the following equation A and B represent the two sets of words to be compared:

$$\text{Jaccard}(A, B) = \frac{A \cap B}{A \cup B} \tag{9}$$

Now we explain how the CONTEXT method finds typically co-occurring words for each of the idiomatic and literal meanings of an expression. Note that unlike in a supervised setting, here we do not assume access to manually annotated training data. We thus use knowledge of automatically acquired canonical forms to find these.

The CONTEXT method labels usages of an expression in a leave-one-out strategy, where each test token is labelled by using the other tokens as noisy training (gold-standard) data. Specifically, to provide gold-standard data for each instance of an expression, we first divide the other instances (of the same expression) into likely-idiomatic and likely-literal groups, where the former group contains usages in canonical form(s) and the latter contains usages in non-canonical form(s). We then pick representative usages from each group by selecting the K instances that are most similar to the instance being labelled (the test token) according to the Jaccard similarity score.

Recall that we assume canonical form(s) are predictive of the idiomatic usages and non-canonical form(s) are indicative of the literal usages of an expression. We thus expect the co-occurrence sets of the selected canonical and non-canonical instances to reflect the idiomatic and literal meanings of the expression, respectively. We take the average similarity of the test token to the K nearest canonical instances (likely idiomatic) and the K nearest non-canonical instances (likely literal), and label the test token accordingly.[19] In the event that there are less than K canonical or non-canonical form usages of an expression, we take the average similarity over however many instances there are of this form. If we have no instances of one of these forms, we classify each token as idiomatic, the label we expect to be more frequent.

7. VNIC Token Identification: Evaluation

To evaluate the performance of our proposed token identification methods, we use each in a classification task, in which the method indicates for each instance of a given expression whether it has an idiomatic or a literal interpretation. Section 7.1 explains the details of our experimental setup. Section 7.2 then presents the experimental results as well as some discussion and analysis.

7.1 Experimental Setup

7.1.1 Experimental Expressions and Annotation. In our token classification experiments, we use a subset of the 180 idiomatic expressions in the development and test data sets used in the type-based experiments of Section 4. From the original 180 expressions, we discard those whose frequency in the BNC is lower than 20, to increase the likelihood that there are both literal and idiomatic usages of each expression. We also discard any

18 It is possible to incorporate extra knowledge sources, such as WordNet, for measuring similarity between two sets of words. However, our intention is to focus on purely unsupervised, knowledge-lean approaches.

19 We also tried using the average similarity of the test token to all instances in each group. However, we found that focusing on the most similar instances from each group performs better.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

expression that is not from the two dictionaries ODCIE and CCID (see Section 4.1.2 for more details on the original data sets). This process results in the selection of 60 candidate verb–noun pairs.

For each of the selected pairs, 100 sentences containing its usage were randomly extracted from the automatically parsed BNC, using the method described in Section 4.1.1. For a pair which occurs less than 100 times in the BNC, all of its usages were extracted. Two judges were asked to independently label each use of each candidate expression as literal, idiomatic, or unknown. When annotating a token, the judges had access to only the sentence in which it occurred, and not the surrounding sentences. If this context was insufficient to determine the class of the expression, the judge assigned the unknown label. In an effort to assure high agreement between the judges' annotations, the judges were also provided with the dictionary definitions of the idiomatic meanings of the expressions.

Idiomaticity is not a binary property; rather it is known to fall on a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic. The human annotators were required to pick the label, literal or idiomatic, that best fit the usage in their judgment; they were not to use the unknown label for intermediate cases. Figurative extensions of literal meanings were classified as literal if their overall meaning was judged to be fairly transparent, as in *You turn right when we* ***hit the road*** *at the end of this track* (taken from the BNC). Sometimes an idiomatic usage, such as *have word* in *At the moment they only* ***had*** *the* ***word*** *of Nicola's husband for what had happened* (also taken from the BNC), is somewhat directly related to its literal meaning, which is not the case for more semantically opaque idioms such as *hit the roof*. This sentence was classified as idiomatic because the idiomatic meaning is much more salient than the literal meaning.

First, our primary judge, a native English speaker and an author of this paper, annotated each use of each candidate expression. Based on this judge's annotations, we removed the 25 expressions with fewer than 5 instances of either of their literal or idiomatic meanings, leaving 28 expressions.[20] (We will revisit the 25 removed expressions in Section 7.2.4.) The remaining expressions were then split into development (DEV) and test (TEST) sets of 14 expressions each. The data was divided such that DEV and TEST would be approximately equal with respect to the frequency of their expressions, as well as their proportion of idiomatic-to-literal usages (according to the primary judge's annotations). At this stage, DEV and TEST contained a total of 813 and 743 tokens, respectively.

Our second judge, also a native English-speaking author of this paper, then annotated DEV and TEST sentences. The observed agreement and unweighted kappa score (Cohen 1960) on TEST were 76% and 0.62, respectively. The judges discussed tokens on which they disagreed to achieve a consensus annotation. Final annotations were generated by removing tokens that received the unknown label as the consensus annotation, leaving DEV and TEST with a total of 573 and 607 tokens, and an average of 41 and 43 tokens per expression, respectively. Table 5 shows the DEV and the TEST verb–noun pairs used in our experiments. The table also contains information on the number of tokens considered for each pair, as well as the percentage of its usages which are idiomatic.

20 From the original set of 60 expressions, seven were excluded because our primary annotator did not provide any annotations for them. These include *catch one's breath*, *cut one's losses*, and *push one's luck* (for which our annotator did not have access to a literal interpretation); and *blow one's (own) horn*, *pull one's hair*, *give a lift*, and *get the bird* (for which our annotator did not have access to an idiomatic meaning).

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 5
Experimental DEV and TEST verb–noun pairs, their token frequency (FRQ), and the percentage of their usages that are idiomatic (%IDM), ordered in decreasing %IDM.

DEV			TEST		
verb–noun	FRQ	%IDM	verb–noun	FRQ	%IDM
find foot	52	90	*have word*	89	90
make face	30	90	*lose thread*	20	90
get nod	26	89	*get sack*	50	86
pull weight	33	82	*make mark*	85	85
kick heel	38	79	*cut figure*	43	84
hit road	31	77	*pull punch*	22	82
take heart	79	73	*blow top*	28	82
pull plug	65	69	*make scene*	48	58
blow trumpet	29	66	*make hay*	17	53
hit roof	17	65	*get wind*	29	45
lose head	38	55	*make hit*	14	36
make pile	25	32	*blow whistle*	78	35
pull leg	51	22	*hold fire*	23	30
see star	61	8	*hit wall*	61	11

7.1.2 Baselines, Parameters, and Performance Measures. We compare the performance of our proposed methods, CFORM and CONTEXT, with the baseline of always predicting an idiomatic interpretation, the most frequent meaning in our development data. We also compare the unsupervised methods against a supervised method, SUP, which is similar to CONTEXT, except that it forms the idiomatic and literal co-occurrence sets from manually annotated data (instead of automatically labelled data using canonical forms). Like CONTEXT, SUP also classifies tokens in a leave-one-out methodology using the K idiomatic and literal instances which are most similar to a test token. For both CONTEXT and SUP, we set the value of K (the number of similar instances used as gold-standard) to 5, since experiments on DEV indicated that performance did not vary substantially using a range of values of K.

For all methods, we report the accuracy macro-averaged over all expressions in TEST. We use the individual accuracies (accuracies for the individual expressions) to perform t-tests for verifying whether different methods have significantly different performance. To further analyze the performance of the methods, we also report their recall and precision on identifying usages from each of the idiomatic and literal classes.

7.2 Experimental Results and Analysis

We first discuss the overall performance of our proposed unsupervised methods in Section 7.2.1. Results reported in Section 7.2.1 are on TEST (results on DEV have similar trends, unless noted otherwise). Next, we look into the performance of our methods on expressions with different proportions of idiomatic-to-literal usages in Section 7.2.2, which presents results on TEST and DEV combined, as explained subsequently. Section 7.2.3 provides an analysis of the errors made because of using canonical forms, and identifies some possible directions for future work. In Section 7.2.4, we present results on a new data set containing expressions with highly skewed proportion of idiomatic-to-literal usages.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 6
Macro-averaged accuracy (%*Acc*) and relative error rate reduction (%*ERR*) on TEST expressions.

Method		%*Acc*	(%*ERR*)
Baseline		61.9	
Unsupervised	CONTEXT	65.8	(10.2)
	CFORM	**72.4**	(**27.6**)
Supervised	SUP	82.7	(54.6)

7.2.1 Overall Performance. Table 6 shows the macro-averaged accuracy on TEST of our two unsupervised methods, as well as that of the baseline and the supervised method for comparison. The best unsupervised performance is indicated in boldface.

As the table shows, both of our unsupervised methods as well as the supervised method outperform the baseline, confirming that the canonical forms of an expression, and local context, are both informative in distinguishing literal and idiomatic instances of the expression.[21] Moreover, CFORM outperforms CONTEXT (difference is marginally significant at $p < .06$), which is somewhat unexpected, as CONTEXT was proposed as an improvement over CFORM in that it combines contextual information along with the syntactic information provided by CFORM. We return to these results later (Section 7.2.3) to offer some reasons as to why this might be the case. However, the results using CFORM confirm our hypothesis that canonical forms—which reflect the overall behavior of a verb+noun type—are strongly informative about the class of a token. Importantly, this is the case even though the canonical forms that we use are imperfect knowledge obtained automatically through an unsupervised method.

Comparing CFORM with SUP, we observe that even though on average the latter outperforms the former, the difference is not statistically significant ($p > .1$). A close look at the performance of these methods on the individual expressions reveals that neither consistently outperforms the other on all (or even most) expressions. Moreover, as we will see in Section 7.2.2, SUP seems to gain most of its advantage over CFORM on expressions with a low proportion of idiomatic usages, for which canonical forms tend to have less predictive value (see Section 7.2.3 for details).

Recall that both CONTEXT and SUP label each token by comparing its local context to those of its *K* nearest "idiomatic" and its *K* nearest "literal" usages. The difference is that CONTEXT uses noisy (automatically) labelled data to identify these nearest usages for each token, whereas SUP uses manually labelled data. One possible direction for future work is thus to investigate whether providing substantially larger amounts of data alleviates the effect of noise, as is often found to be the case by researchers in the field.

7.2.2 Performance Based on Class Distribution. Recall from Section 6 that both of our unsupervised techniques for token identification depend on how accurately the canonical forms of an expression can be acquired. The canonical form acquisition technique which we use here works well if the idiomatic meaning of an expression is sufficiently frequent compared to its literal usage. In this section, we thus examine the performance of the

21 Performing a paired t-test, we find that the difference between the baseline and CFORM is marginally significant, $p < .06$, whereas the difference between baseline and CONTEXT is not statistically significant. The difference between the baseline and SUP is significant at $p < .01$. The trend on DEV is somewhat similar: baseline and CFORM are significantly different at $p < .05$; SUP is marginally different from baseline at $p < .06$.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 7
Macro-averaged accuracy (%*Acc*) and relative error rate reduction (%*ERR*) on the 28 expressions in DT (DEV and TEST combined), divided according to the proportion of idiomatic-to-literal usages (*high* and *low*).

		$DT_{I_{high}}$		$DT_{I_{low}}$	
Method		%*Acc*	(%*ERR*)	%*Acc*	(%*ERR*)
Baseline		81.4		35.0	
Unsupervised	CONTEXT	80.6	(−4.3)	44.6	(14.8)
	CFORM	**84.7**	(17.7)	53.4	(28.3)
Supervised	SUP	84.4	(16.1)	**76.8**	(64.3)

token identification methods for expressions with different proportions of idiomatic-to-literal usages.

We merge DEV and TEST (referring to the new set as DT), and then divide the resulting set of 28 expressions according to their proportion of idiomatic-to-literal usages (as determined by the human annotations) as follows.[22] Looking at the proportion of idiomatic usages of our expressions in Table 5, we can see that there are gaps between 55% and 65% in DEV, and between 58% and 82% in TEST, in terms of proportion of idiomatic usages. The value of 65% thus serves as a natural lower bound for dominant idiomatic usage, and the value of 58% as a natural upper bound for non-dominant idiomatic usage. We therefore split DT into two sets: $DT_{I_{high}}$ contains 17 expressions with 65–90% of their usages being idiomatic (i.e., their idiomatic usage is dominant), whereas $DT_{I_{low}}$ contains 11 expressions with 8–58% of their occurrences being idiomatic (i.e., their idiomatic usage is not dominant).

Table 7 shows the average accuracy of all the methods on these two groups of expressions, with the best performance on each group shown in boldface. We first look at the performance of our methods on $DT_{I_{high}}$. On these expressions, CFORM outperforms both the baseline (difference is not statistically significant) and CONTEXT (difference is statistically significant at $p < .05$). CFORM also has a comparable performance to the supervised method, reinforcing that for these expressions accurate canonical forms can be acquired and that such knowledge can be used with high confidence for distinguishing idiomatic and literal usages in context.

We now look into the performance on expressions in $DT_{I_{low}}$. On these, both CFORM and CONTEXT outperform the baseline, showing that even for expressions whose idiomatic meaning is not dominant, automatically acquired canonical forms can help with their token classification. Nonetheless, both these methods perform substantially worse than the supervised method, reinforcing that the automatically acquired canonical forms are noisier, and hence less predictive, than they are for expressions in $DT_{I_{high}}$.

The poor performance of the unsupervised methods on expressions in $DT_{I_{low}}$ (compared to the supervised performance) is likely to be mostly due to the less predictive canonical forms extracted for these expressions. In general, we can conclude that when canonical forms can be extracted with a high accuracy, the performance of the CFORM method is comparable to that of a supervised method. One possible way of improving the performance of unsupervised methods is thus to develop more accurate techniques for the automatic acquisition of canonical forms.

22 We combine the two sets in order to have a sufficient number of expressions in each group after division.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 8
Confusion matrix for CFORM on expression *blow trumpet*. idm = idiomatic class; lit = literal class; *tp* = true positive; *fp* = false positive; *fn* = false negative; *tn* = true negative.

		True Class	
		idm	lit
Predicted Class	idm	17 = *tp*	6 = *fp*
	lit	2 = *fn*	4 = *tn*

Table 9
Formulas for calculating Sens and PPV (recall and precision for the idiomatic class), and Spec and NPV (recall and precision for the literal class) from a confusion matrix.

	recall (R)			precision (P)		
idm	Sens	=	$\frac{tp}{tp+fn}$	PPV	=	$\frac{tp}{tp+fp}$
lit	Spec	=	$\frac{tn}{tn+fp}$	NPV	=	$\frac{tn}{tn+fn}$

Accuracy is often not a sufficient measure for the evaluation of a binary (two-class) classifier, especially when the number of items in the two classes (here, idiomatic and literal) differ. Instead, one can have a closer look at the performance of a classifier by examining its confusion matrix, which compares the labels predicted by the classifier for each item with its true label. As an example, the confusion matrix of the CFORM method for the expression *blow trumpet* is given in Table 8.

Note that the choice of idiomatic as the positive class (and literal as the negative class) is arbitrary; however, because our ultimate goal is to identify idiomatic usages, there is a natural reason for this choice. To summarize a confusion matrix, four standard measures are often used, which are calculated from the cells in the matrix. The measures are sensitivity (Sens), positive predictive value (PPV), specificity (Spec), and negative predictive value (NPV), and are calculated as in Table 9. As stated in the table, Sens and PPV are equivalents of recall and precision for the positive (idiomatic) class, also referred to as R_{idm} and P_{idm} later in the article. Similarly, Spec and NPV are equivalents of recall and precision for the negative (literal) class, also referred to as R_{lit} and P_{lit}.[23]

Table 10 gives the trimmed mean values of these four performance measures over expressions in $\text{DT}_{\text{I}_{high}}$ and $\text{DT}_{\text{I}_{low}}$ for the baseline, the two unsupervised methods, and the supervised method.[24] (The performance measures on individual expressions are given in Tables 12, 13, and 14 in the Appendix.) Table 10 shows that, as expected, the baseline has very high Sens (100% recall on identifying idiomatic usages), but very low Spec (0%

23 We mainly refer to these measures using their standard names in the literature: Sens, PPV, Spec, and NPV. Alongside the standard names, we use the more expressive names R_{idm}, P_{idm}, R_{lit}, and P_{lit}, to remind the reader about the semantics of the measures.

24 When averaging interdependent measures, such as precision and recall, one needs to make sure that the observed trend in the averages is consistent with that in the individual values. **Trimmed mean** is a standard statistic used in such cases, which is equivalent to the mean after discarding a percentage (often between 5 and 25) of the sample data at the high and low ends. Here, we report a 14%-trimmed mean, which involves removing two data points from each end. The analysis presented here is based on the trimmed means, as well as the individual values of the performance measures.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 10
Detailed classification performance of all methods over $\text{DT}_{\text{I}_{high}}$ and $\text{DT}_{\text{I}_{low}}$. Performance is given using four measures: Sens or R_{idm}, PPV or P_{idm}, Spec or R_{lit}, and NPV or P_{lit}, macro-averaged using 14%-trimmed mean.

Data Set	Method	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
$\text{DT}_{\text{I}_{high}}$	Baseline	1.00	.82	0.00	0.00
	CONTEXT	.97	.84	.11	.18
	CFORM	.95	.92	.61	.71
	SUP	.99	.86	.22	.53
Data Set	Method	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
$\text{DT}_{\text{I}_{low}}$	Baseline	1.00	.36	0.00	0.00
	CONTEXT	.89	.37	.22	.63
	CFORM	.86	.43	.36	.86
	SUP	.44	.62	.88	.80

recall on identifying literal usages). We thus expect a well-performing method to have lower Sens than the baseline, but higher Spec and also higher PPV and NPV (i.e., higher precision on both idiomatic and literal usages).

Looking at performance on $\text{DT}_{\text{I}_{high}}$, we find that all three methods have reasonably high Sens and PPV, revealing that the methods are good at labeling idiomatic usages. Performance on literal usages, however, differs across the three methods. CONTEXT has very low Spec and NPV, showing that it tends to label most tokens—including the literal ones—as idiomatic. A close look at the performance of this method on the individual expressions also confirms this tendency: on many expressions (10 out of 17) the Spec and NPV of CONTEXT are both zero (see Table 13 in the Appendix). As we will see in Section 7.2.3, this tendency is partly due to the distribution of the idiomatic and literal usages in canonical and non-canonical forms; because literal usages can also appear in a canonical form, for many expressions there are often not many non-canonical form instances. (Recall that, for training, CONTEXT uses instances in canonical form as being idiomatic and those in non-canonical form as being literal.) Thus, in many cases, it is a priori more likely that a token is more similar to the K most similar canonical form instances. Interestingly, CFORM is the method with the highest Spec and NPV, even higher than those of the supervised method. Nonetheless, even CFORM is overall much better at identifying idiomatic tokens than literal ones (see Section 7.2.3 for more discussion on this).

We now turn to performance on $\text{DT}_{\text{I}_{low}}$. CFORM has a high Sens, but a low PPV, indicating that most idiomatic usages are identified correctly, but many literal usages are also misclassified as idiomatic (hence a low Spec). CONTEXT shows the same trend as CFORM, though overall it has poorer performance. Performance of SUP varies across the expressions in this group: SUP is very good at identifying literal usages of these expressions (high Spec and NPV for all expressions). Nonetheless, SUP has a low recall in identifying idiomatic usages (low Sens) for many of these expressions.

7.2.3 Discussion and Error Analysis. In this section, we examine two main issues. First, we look into the plausibility of our original assumptions regarding the predictive value of canonical forms (and non-canonical forms). Second, we investigate the appropriateness of our automatically extracted canonical forms.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

To learn more about the predictive value of canonical forms, we examine the performance of CFORM on the 28 expressions under study. More specifically, we look at the values of Sens, PPV, Spec, and NPV on these expressions, as shown in Table 12 in the Appendix. On expressions in $\text{DT}_{\text{I}_{high}}$, CFORM has both high Sens and high PPV. The formulas in Table 9 indicate that if both Sens and PPV are high, then $tp \gg fn$ and $tp \gg fp$. Thus, most idiomatic usages of expressions in $\text{DT}_{\text{I}_{high}}$ appear in a canonical form, and most usages in a canonical form are idiomatic. The values of Spec and NPV on the same expressions are in general lower (compared to Sens and PPV), showing that *tn* is not much higher than *fp* or *fn*.

On expressions in $\text{DT}_{\text{I}_{low}}$, CFORM generally has high Sens but low-to-medium PPV. This indicates that for these expressions, most idiomatic usages appear in a canonical form, but not all usages in a canonical form are idiomatic. On these expressions, CFORM has generally high NPV, but mostly low Spec. These indicate that $tn \gg fn$, that is, most usages in a non-canonical form are literal, and that *tn* is often lower than *fp*, that is, many literal usages also appear in a canonical form. For example, almost all usages of *hit wall* in a non-canonical form are literal, but most of its literal usages appear in a canonical form.

Generally, it seems that, as we expected, literal usages are less restricted in terms of the syntactic form they appear in; they can appear in both canonical form(s) and in non-canonical form(s). For an expression with a low proportion of literal usages, we can thus acquire canonical forms that are both accurate and have high predictive value for identifying idiomatic usages in context. On the contrary, for expressions with a relatively high proportion of literal usages, automatically acquired canonical forms are less accurate and also have low predictive value (i.e., they are not specific to idiomatic usages). We expected that using contextual information would help in such cases. However, our CONTEXT method relies on noisy training data automatically labelled using information about canonical forms. Given these findings, it is not surprising that this method performs substantially worse than a corresponding supervised method that uses similar contextual information, but manually labelled training data. It remains to be tested in the future whether providing more noisy data will help. Another possible future direction is to develop context methods that can better exploit noisy labelled data.

Now we look at a few cases where our automatically extracted canonical forms are not sufficiently accurate. For a verb+noun such as *make pile* (i.e., *make a pile of money*), we correctly identify only some of the canonical forms. The automatically determined canonical forms for *make pile* are *make a pile* and *make piles*. However, we find that idiomatic usages of this expression are sometimes of the form *make one's pile*. Furthermore, we find that the frequency of this form is much higher than that of the non-canonical forms, and not substantially lower than the frequency cut-off for selection as a canonical form. This indicates that our heuristic for selecting patterns as canonical forms could be fine-tuned to yield an improvement in performance.

For the expression *pull plug*, we identify its canonical form as *pull the plug*, but find a mixture of literal and idiomatic usages in this form. However, many of the literal usages are verb-particle constructions using *out* (*pull the plug out*), while many of the idiomatic usages occur with a prepositional phrase headed by *on* (*pull the plug on*). This indicates that incorporating information about particles and prepositions could improve the quality of the canonical forms. Other syntactic categories, such as adjectives, may also be informative in determining canonical forms for expressions which are typically used idiomatically with words of a particular syntactic category, as in *blow one's own trumpet*.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 11
Macro-averaged accuracy (%*Acc*) and relative error rate reduction (%*ERR*) on the 23 expressions in SKEWED-IDM and on the 37 expressions in the combination of TEST and SKEWED-IDM (ALL).

		SKEWED-IDM		ALL	
Method		%*Acc*	(%*ERR*)	%*Acc*	(%*ERR*)
Baseline		97.9		84.3	
Unsupervised	CONTEXT	94.2	(−176.2)	83.3	(−6.4)
	CFORM	86.7	(−533.3)	81.3	(−19.1)
Supervised	SUP	97.9	(0.0)	92.1	(49.7)

7.2.4 Performance on Expressions with Skewed Distribution. Recall from Section 7.1.1 that, from the original set of 60 candidate expressions, we excluded those that had fewer than 5 instances of either of their literal or idiomatic meanings. It is nonetheless important to see how well our methods perform on such expressions. In this section, we thus report the performance of our measures on the set of 23 expressions with mostly idiomatic usages, referred to as SKEWED-IDM. Table 11 presents the macro-averaged accuracy of our methods on these expressions. This table also shows the accuracy on all unseen test expressions, that is, the combination of SKEWED-IDM and TEST, referred to as ALL, for comparison.[25]

On SKEWED-IDM, the supervised method performs as well as the baseline, whereas both unsupervised methods perform worse.[26] Note that for 19 out of the 23 expressions in SKEWED-IDM, all instances are idiomatic, and the baseline accuracy is thus 100%. On these, SUP also has 100% accuracy because no literal instances are available, and thus SUP labels every token as idiomatic (same as the baseline). As for the unsupervised methods, we can see that, unlike on TEST, the CONTEXT method outperforms CFORM (the difference is statistically significant at $p < .001$). We saw previously that CONTEXT tends to label usages as idiomatic. This bias might be partially responsible for the better performance of CONTEXT on this data set. Moreover, we find that many of these expressions tend to appear in a highly frequent canonical form, but also in less frequent syntactic forms which we (perhaps incorrectly) consider as non-canonical forms. When considering the performance on all unseen test expressions (ALL), neither unsupervised method performs as well as the baseline, but the supervised method offers a substantial improvement over the baseline.[27]

Our annotators pointed out that for many of the expressions in SKEWED-IDM, either a literal interpretation was almost impossible (as for *catch one's imagination*), or extremely implausible (as for *kick the habit*). Hence, the annotators could predict beforehand that the expression would be mainly used with an idiomatic meaning. A semi-supervised approach that combines expert human knowledge with automatically extracted corpus-drawn information can thus be beneficial for the task of identifying

25 The results obtained on the two excluded expressions which are predominantly used literally in terms of percent accuracy using the various methods are as follows. Baseline: 4.2, Unsupervised CONTEXT: 6.5, Unsupervised CFORM: 16.2, Supervised: 43.5. However, because there are only two such expressions, it is difficult to draw conclusions from these results, and we do not further consider these expressions.

26 According to a paired t-test, on SKEWED-IDM, all the observed differences are statistically significant at $p < .05$.

27 According to a paired t-test, on ALL, the differences between the supervised method and the three other methods are statistically significant at $p < .01$; none of the other differences are statistically significant.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

idiomatic expressions in context. A human expert (e.g., a lexicographer) could first filter out expressions for which a literal interpretation is highly unlikely. For the rest of the expressions, a simple unsupervised method such as CFORM—that relies only on automatically extracted information—can be used with reasonable accuracy.

8. Related Work

8.1 Type-Based Recognition of Idioms and Other Multiword Expressions

Our work relates to previous studies on determining the compositionality (the inverse of idiomaticity) of idioms and other multiword expressions (MWEs). Most previous work on the compositionality of MWEs either treats them as collocations (Smadja 1993), or examines the distributional similarity between the expression and its constituents (Baldwin et al. 2003; Bannard, Baldwin, and Lascarides 2003; McCarthy, Keller, and Carroll 2003). Others have identified MWEs by looking into specific linguistic cues, such as the lexical fixedness of non-compositional MWEs (Lin 1999; Wermter and Hahn 2005), or the lexical flexibility of productive noun compounds (Lapata and Lascarides 2003). Venkatapathy and Joshi (2005) combine aspects of this work, by incorporating lexical fixedness, distributional similarity, and collocation-based measures into a set of features which are used to rank verb+noun combinations according to their compositionality. Our work differs from such studies in that it considers various kinds of fixedness as surface behaviors that are tightly related to the underlying semantic idiosyncrasy (idiomaticity) of expressions. Accordingly, we propose novel methods for measuring the degree of lexical, syntactic, and overall fixedness of verb+noun combinations, and use these as indirect ways of measuring degree of idiomaticity.

Earlier research on the lexical encoding of idiom types mainly relied on the existence of human annotations, especially for detecting which syntactic variations (e.g., passivization) an idiom can undergo (Odijk 2004; Villavicencio et al. 2004). Evert, Heid, and Spranger (2004) and Ritz and Heid (2006) propose methods for automatically determining morphosyntactic preferences of idiomatic expressions. However, they treat individual morphosyntactic markers (e.g., the number of the noun in a verb+noun combination) as independent features, and rely mainly on the relative frequency of each possible value for a feature (e.g., plural for number) as an indicator of a preference for that value. If the relative frequency of a particular value of a feature for a given combination (or the lower bound of the confidence interval, in the case of Evert, Heid, and Spranger's approach) is higher than a certain threshold, then the expression is said to have a preference for that value. These studies recognize that morphosyntactic preferences can be employed as clues to the identification of idiomatic combinations; however, none proposes a systematic approach for such a task. Moreover, only subjective evaluations of the proposed methods are presented.

Others have also drawn on the notion of syntactic fixedness for the detection of idioms and other MWEs. Widdows and Dorow (2005), for example, look into the fixedness of a highly constrained type of idiom, namely, those of the form "X **conj** X" where X is a noun or an adjective, and **conj** is a conjunction such as *and*, *or*, *but*. Smadja (1993) also notes the importance of syntactic fixedness in identifying strongly associated multiword sequences, including collocations and idioms. Nonetheless, in both these studies, the notion of syntactic fixedness is limited to the relative position of words within the sequence. Such a general notion of fixedness does not take into account some of the important syntactic properties of idioms (e.g., the choice of the determiner), and hence cannot distinguish among different subtypes of MWEs which may differ on such

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

grounds. Our syntactic fixedness measure looks into a set of linguistically informed patterns associated with a coherent, though large, class of idiomatic expressions. Results presented in this article show that the fixedness measures can successfully separate idioms from literal phrases. Corpus analysis of the measures proves that they can also be used to distinguish idioms from other MWEs, such as light verb constructions and collocations (Fazly and Stevenson 2007; Fazly and Stevenson, to appear). Bannard (2007) proposes an extension of our syntactic fixedness measure—which first appeared in Fazly and Stevenson (2006)—where he uses different prior distributions for different syntactic variations.

Work on the identification of MWE types has also looked at evidence from another language. For example, Melamed (1997a) assumes that non-compositional compounds (NCCs) are usually not translated word-for-word to another language. He thus proposes to discover NCCs by maximizing the information-theoretic predictive value of a translation model between two languages. The sample extracted NCCs reveal an important drawback of the proposed method: It relies on a translation model only, without taking into account any prior linguistic knowledge about possible NCCs within a language. Nonetheless, such a technique is capable of identifying many NCCs that are relevant for a translation task. Villada Moirón and Tiedemann (2006) propose measures for distinguishing idiomatic expressions from literal ones (in Dutch), by examining their automatically generated translations into a second language, such as English or Spanish. Their approach is based on the assumptions that idiomatic expressions tend to have fewer predictable translations and fewer compositional meanings, compared to the literal ones. The first property is measured as the diversity in the translations for the expression, estimated using an entropy-based measure proposed by Melamed (1997b). The non-compositionality of an expression is measured as the overlap between the meaning of an expression (i.e., its translations) and those of its component words.

General approaches (such as those explained in the previous paragraph) may be more easily extended to different domains and languages. Our measures incorporate language-specific information about idiomatic expressions, thus extra work may be required to extend and apply them to other languages and other expressions. (Though see Van de Cruys and Villada Moirón [2007] for an extension of our measures to Dutch idioms of the form verb plus prepositional phrase.) Nonetheless, because our measures capture deep linguistic information, they are also expected to acquire more detailed knowledge—for example, they can be used for identifying other classes of MWEs (Fazly and Stevenson 2007).

8.2 Token-Based Identification of Idioms and Other Multiword Expressions

A handful of studies have focused on identifying idiomatic and non-idiomatic usages (tokens) of words or MWEs. Birke and Sarkar (2006) propose a minimally supervised algorithm for distinguishing between literal and non-literal usages of verbs in context. Their algorithm uses seed sets of literal and non-literal usages that are automatically extracted from online resources such as WordNet. The similarity between the context of a target token and that of each seed set determines the class of the token. The approach is general in that it uses a slightly modified version of an existing word sense disambiguation algorithm. This is both an advantage and a drawback: The algorithm can be easily extended to other parts of speech and other languages; however, such a general method ignores the specific properties of non-literal (metaphorical and/or idiomatic) language. Similarly, the supervised token classification method of Katz and Giesbrecht (2006) relies primarily on the local context of a token, and fails to exploit specific linguistic

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

properties of non-literal language. Our results suggest that such properties are often more informative than the local context, in determining the class of an MWE token.

The supervised classifier of Patrick and Fletcher (2005) distinguishes between compositional and non-compositional usages of English verb-particle constructions. Their classifier incorporates linguistically motivated features, such as the degree of separation between the verb and particle. Here, we focus on a different class of English MWEs, namely, the class of idiomatic verb+noun combinations. Moreover, by making a more direct use of their syntactic behavior, we develop unsupervised token classification methods that perform well. The unsupervised token classifier of Hashimoto, Sato, and Utsuro (2006) uses manually encoded information about allowable and non-allowable syntactic transformations of Japanese idioms, which are roughly equivalent to our notions of canonical and non-canonical forms. The rule-based classifier of Uchiyama, Baldwin, and Ishizaki (2005) incorporates syntactic information about Japanese compound verbs (JCVs), a type of MWE composed of two verbs. In both cases, although the classifiers incorporate syntactic information about MWEs, their manual development limits the scalability of the approaches.

Uchiyama, Baldwin, and Ishizaki (2005) also propose a statistical token classification method for JCVs. This method is similar to ours, in that it also uses type-based knowledge to determine the class of each token in context. However, their method is supervised, whereas our methods are unsupervised. Moreover, Uchiyama, Baldwin, and Ishizaki only evaluate their methods on a set of JCVs that are mostly monosemous. Our main focus here is on MWEs that are harder to disambiguate, that is, those that have two clear idiomatic and literal meanings, and that are frequently used with either meaning.

9. Conclusions

The significance of the role idioms play in language has long been recognized; however, due to their peculiar behavior, they have been mostly overlooked by researchers in computational linguistics. In this work, we focus on a broadly documented and cross-linguistically frequent class of idiomatic MWEs: those that involve the combination of a verb and a noun in its direct object position, which we refer to as verb+noun idiomatic combinations or VNICs. Although a great deal of research has focused on non-compositionality of MWEs, less attention has been paid to other properties relevant to their semantic idiosyncrasy, such as lexical and syntactic fixedness. Drawing on such properties, we have developed techniques for the automatic recognition of VNIC types, as well as methods for their token identification in context.

We propose techniques for the automatic acquisition and encoding of knowledge about the lexicosyntactic behavior of idiomatic combinations. More specifically, we propose novel statistical measures that quantify the degree of lexical, syntactic, and overall fixedness of a verb+noun combination. We demonstrate that these measures can be successfully applied to the task of automatically distinguishing idiomatic expressions (types) from non-idiomatic ones. Our results show that the syntactic and overall fixedness measures substantially outperform existing measures of collocation extraction, even when they incorporate some syntactic information. We put forward an unsupervised means for automatically discovering the set of syntactic variations that are preferred by a VNIC type (its canonical forms) and that should be included in its lexical representation. In addition, we show that the canonical form extraction method can effectively be used in identifying idiomatic and literal usages (tokens) of an expression in context.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

We have annotated a total of 2,465 tokens for 51 VNIC types according to whether they are a literal or idiomatic usage. We found that for 28 expressions (1,180 tokens), approximately 40% of the usages were literal. For the remaining 23 expressions (1,285 tokens), almost all usages were idiomatic. These figures indicate that automatically determining whether a particular instance of an expression is used idiomatically or literally is of great importance for NLP applications. We have proposed two unsupervised methods that perform such a task.

Our proposed methods incorporate automatically acquired knowledge about the overall syntactic behavior of a VNIC type, in order to do token classification. More specifically, our methods draw on the syntactic fixedness of VNICs—a property which has been largely ignored in previous studies of MWE tokens. Our results confirm the usefulness of this property as incorporated into our methods. On the 23 expressions whose usages are predominantly idiomatic, because the baseline is very high none of the methods outperform it. Nonetheless, as pointed out by our human annotators, for many of these expressions it can be predicted beforehand that they are mainly idiomatic and that a literal interpretation is impossible or highly implausible. On the 28 expressions with frequent literal usages, all our methods outperform the baseline of always predicting the most dominant class (idiomatic). Moreover, on these, the accuracy of our best unsupervised method is not substantially lower than the accuracy of a standard supervised approach.

Appendix: Performance on the Individual Expressions

This Appendix contains the values of the four performance measures, Sens, PPV, Spec, and NPV, for our two unsupervised methods (i.e., CFORM and CONTEXT) as well as for the supervised method, SUP, on individual expressions in $\text{DT}_{\text{I}_{high}}$ and $\text{DT}_{\text{I}_{low}}$. Expressions (verb–noun pairs) in each data set are ordered alphabetically.

Table 12
Performance of CFORM on individual expressions in $\text{DT}_{\text{I}_{high}}$ and $\text{DT}_{\text{I}_{low}}$.

Data Set	verb–noun	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
$\text{DT}_{\text{I}_{high}}$	*blow top*	1.00	0.92	0.60	1.00
	blow trumpet	0.89	0.89	0.80	0.80
	cut figure	0.97	0.97	0.86	0.86
	find foot	0.98	0.92	0.20	0.50
	get nod	0.96	1.00	1.00	0.75
	get sack	1.00	0.96	0.71	1.00
	have word	0.56	0.96	0.78	0.17
	hit road	1.00	0.80	0.14	1.00
	hit roof	1.00	0.65	0.00	0.00
	kick heel	1.00	0.81	0.12	1.00
	lose thread	0.94	0.94	0.50	0.50
	make face	0.74	0.95	0.67	0.22
	make mark	0.85	1.00	1.00	0.54
	pull plug	0.89	0.77	0.40	0.62
	pull punch	0.83	0.94	0.75	0.50
	pull weight	1.00	0.93	0.67	1.00
	take heart	1.00	0.97	0.88	1.00

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 12
(*continued*)

Data Set	verb–noun	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
	blow whistle	0.93	0.44	0.37	0.90
	get wind	0.85	0.73	0.75	0.86
	hit wall	0.86	0.11	0.09	0.83
	hold fire	1.00	0.37	0.25	1.00
	lose head	0.76	0.62	0.41	0.58
$\text{DT}_{\text{I}_{low}}$	*make hay*	1.00	0.56	0.12	1.00
	make hit	1.00	0.71	0.78	1.00
	make pile	0.25	0.14	0.29	0.45
	make scene	0.82	0.68	0.45	0.64
	pull leg	0.64	0.23	0.40	0.80
	see star	0.80	0.10	0.38	0.95

Table 13
Performance of CONTEXT on individual expressions in $\text{DT}_{\text{I}_{high}}$ and $\text{DT}_{\text{I}_{low}}$.

Data Set	verb–noun	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
	blow top	1.00	0.85	0.20	1.00
	blow trumpet	0.89	0.74	0.40	0.67
	cut figure	1.00	0.84	0.00	0.00
	find foot	1.00	0.90	0.00	0.00
	get nod	1.00	0.88	0.00	0.00
	get sack	1.00	0.86	0.00	0.00
	have word	0.70	0.95	0.67	0.20
	hit road	1.00	0.77	0.00	0.00
$\text{DT}_{\text{I}_{high}}$	*hit roof*	1.00	0.65	0.00	0.00
	kick heel	0.97	0.78	0.00	0.00
	lose thread	1.00	0.90	0.00	0.00
	make face	0.85	0.88	0.00	0.00
	make mark	1.00	0.91	0.46	1.00
	pull plug	0.96	0.69	0.05	0.33
	pull punch	0.94	0.89	0.50	0.67
	pull weight	1.00	0.82	0.00	0.00
	take heart	0.90	0.85	0.38	0.50
	blow whistle	0.89	0.36	0.18	0.75
	get wind	0.85	0.65	0.62	0.83
	hit wall	1.00	0.11	0.00	0.00
	hold fire	1.00	0.30	0.00	0.00
	lose head	0.90	0.56	0.12	0.50
$\text{DT}_{\text{I}_{low}}$	*make hay*	0.78	0.50	0.12	0.33
	make hit	0.60	0.38	0.44	0.67
	make pile	0.50	0.25	0.29	0.56
	make scene	0.96	0.66	0.30	0.86
	pull leg	0.82	0.22	0.20	0.80
	see star	1.00	0.12	0.32	1.00

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Table 14
Performance of SUP on individual expressions in $DT_{I_{high}}$ and $DT_{I_{low}}$.

Data Set	verb–noun	Sens (R_{idm})	PPV (P_{idm})	Spec (R_{lit})	NPV (P_{lit})
$DT_{I_{high}}$	*blow top*	1.00	0.85	0.20	1.00
	blow trumpet	0.95	0.72	0.30	0.75
	cut figure	1.00	0.84	0.00	0.00
	find foot	1.00	0.90	0.00	0.00
	get nod	0.91	0.91	0.33	0.33
	get sack	1.00	0.86	0.00	0.00
	have word	1.00	0.90	0.00	0.00
	hit road	1.00	0.80	0.14	1.00
	hit roof	0.82	0.64	0.17	0.33
	kick heel	0.97	0.78	0.00	0.00
	lose thread	1.00	0.95	0.50	1.00
	make face	1.00	0.96	0.67	1.00
	make mark	1.00	0.91	0.46	1.00
	pull plug	0.98	0.90	0.75	0.94
	pull punch	1.00	0.90	0.50	1.00
	pull weight	1.00	0.82	0.00	0.00
	take heart	0.93	0.83	0.25	0.50
$DT_{I_{low}}$	*blow whistle*	0.52	0.78	0.92	0.78
	get wind	0.77	0.71	0.75	0.80
	hit wall	0.00	0.00	1.00	0.89
	hold fire	0.00	0.00	0.88	0.67
	lose head	0.48	0.62	0.65	0.50
	make hay	0.89	0.80	0.75	0.86
	make hit	0.40	1.00	1.00	0.75
	make pile	0.38	0.75	0.94	0.76
	make scene	0.89	0.69	0.45	0.75
	pull leg	0.55	0.75	0.95	0.88
	see star	0.00	0.00	1.00	0.92

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Acknowledgments

This article is an extended and updated combination of two papers that appeared, respectively, in the proceedings of EACL 2006 and the proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions. We wish to thank the anonymous reviewers of those papers for their helpful recommendations. We also thank the anonymous reviewers of this article for their insightful comments which we believe have helped us improve the quality of the work. We are grateful to Eric Joanis for providing us with the NP-head extraction software, and to Afra Alishahi and Vivian Tsang for proofreading the manuscript. Our work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the Ontario Graduate Scholarship program, and the University of Toronto.

References

Abeillé, Anne. 1995. The flexibility of French idioms: A representation with lexicalized Tree Adjoining Grammar. In Everaert et al., editors, *Idioms: Structural and Psychological Perspectives*. LEA, Mahwah, NJ, pages 15–42.

Akimoto, Minoji. 1999. Collocations and idioms in Late Modern English. In L. J. Brinton and M. Akimoto. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. John Benjamins Publishing Company, Amsterdam, pages 207–238.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo.

Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL'07 Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8, Prague.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo.

Birke, Julia and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 329–336, Trento.

Burnard, Lou. 2000. *Reference Guide for the British National Corpus (World Edition)*, second edition. Available at `www.natcorp.ox.ac.uk`.

Cacciari, Cristina. 1993. The place of idioms in a literal and metaphorical world. In C. Cacciari and P. Tabossi, *Idioms: Processing, Structure, and Interpretation*. LEA, Mahwah, NJ, pages 27–53.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. LEA, Mahwah, NJ, pages 115–164.

Claridge, Claudia. 2000. *Multi-word Verbs in Early Modern English: A Corpus-based Study*. Editions Rodopi B. V., Amsterdam.

Clark, Eve V. 1978. Discovering what words can do. *Papers from the Parasession on the Lexicon*, 14:34–57.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL'07 Workshop on a Broader Perspective on Multiword Expressions*, pages 41–48, Prague.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'02)*, pages 1941–47, Las Palmas.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley and Sons, Inc., New York.

Cowie, Anthony P., Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

Dagan, Ido, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word co-occurrence probabilities. In *Proceedings of the 32nd Anuual Meeting of the*

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Association for Computational Linguistics (ACL'94), pages 272–278, Las Cruces, NM.

d'Arcais, Giovanni B. Flores. 1993. The comprehension and semantic interpretation of idioms. In C. Cacciari and P. Tabossi, *Idioms: Processing, Structure, and Interpretation*. LEA, Mahwah, NJ, pages 79–98.

Desbiens, Marguerite Champagne and Mara Simon. 2003. Déterminants et locutions verbales. Manuscript. Available at `www.er.uqam.ca/nobel/scilang/cesla02/mara_margue.pdf`.

Evert, Stefan, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 907–910, Lisbon.

Evert, Stefan and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, pages 188–195, Toulouse.

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 337–344, Trento.

Fazly, Afsaneh and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL'07 Workshop on a Broader Perspective on Multiword Expressions*, pages 9–16, Prague.

Fazly, Afsaneh and Suzanne Stevenson. A distributional account of the semantics of multiword expressions. To appear in the *Italian Journal of Linguistics*.

Fellbaum, Christiane. 1993. The determiner in English idioms. In C. Cacciari and P. Tabossi, *Idioms: Processing, Structure, and Interpretation*. LEA, Mahwah, NJ, pages 271–295.

Fellbaum, Christiane, editor. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fellbaum, Christiane. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In *Proceedings of the KONVENS 2002 Conference*, pages 7–11, Saarbruecken, Germany.

Fellbaum, Christiane. 2007. The ontological loneliness of idioms. In Andrea Schalley and Dietmar Zaefferer, editors, *Ontolinguistics*. Mouton de Gruyter, Berlin, pages 419–434.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*. The Philological Society, Oxford, pages 1–32.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language*, 6:22–42.

Gentner, Dedre and Ilene M. France. 2004. The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In Steven L. Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Kaufmann, San Mateo, CA, pages 343–382.

Gibbs, Raymond W. Jr. 1993. Why idioms are not dead metaphors. In C. Cacciari and P. Tabossi, *Idioms: Processing, Structure, and Interpretation*. LEA, Mahwah, NJ, pages 57–77.

Gibbs, Raymond W. Jr. 1995. Idiomaticity and human cognition. In Everaert et al., editors, *Idioms: Structural and Psychological Perspectives*. LEA, Mahwah, NJ, pages 97–116.

Gibbs, Raymond W. Jr. and Nandini P. Nayak. 1989. Psycholololinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21:100–138.

Gibbs, Raymond W. Jr., Nandini P. Nayak, J. Bolton, and M. Keppel. 1989. Speaker's assumptions about the lexical flexibility of idioms. *Memory and Cognition*, 17:58–68.

Glucksberg, Sam. 1993. Idiom meanings and allusional content. In C. Cacciari and P. Tabossi, *Idioms: Processing, Structure, and Interpretation*. LEA, Mahwah, NJ, pages 3–26.

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

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

Hashimoto, Chikara, Satoshi Sato, and Takehito Utsuro. 2006. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the*

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Association for Computational Linguistics (COLING-ACL'06), pages 353–360, Sydney.
Inkpen, Diana. 2003. *Building a Lexical Knowledge-Base of Near-Synonym Differences*. Ph.D. thesis, University of Toronto.
Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.
Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis. In *Proceedings of the ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney.
Katz, Jerrold J. 1973. Compositionality, idiomaticity, and lexical substitution. In S. Anderson and P. Kiparsky, editors, *A Festschrift for Morris Halle*. Holt, Rinehart and Winston, New York, pages 357–376.
Kearns, Kate. 2002. Light verbs in English. Manuscript. Available at `www.ling.canterbury.ac.nz/people/kearns.html`.
Kirkpatrick, E. M. and C. M. Schwarz, editors. 1982. *Chambers Idioms*. W & R Chambers Ltd, Edinburgh.
Krenn, Brigitte and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL'01 Workshop on Collocations*, pages 39–46, Toulouse.
Kytö, Merja. 1999. Collocational and idiomatic aspects of verbs in Early Modern English. In L. J. Brinton and M. Akimoto. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. John Benjamins Publishing Company, Amsterdam, pages 167–206.
Lapata, Mirella and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 235–242, Budapest.
Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pages 768–774, Montreal.
Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 317–324, College Park, Maryland.
Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.
McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo.
Melamed, I. Dan. 1997a. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, pages 97–108, Providence, RI.
Melamed, I. Dan. 1997b. Measuring semantic entropy. In *Proceedings of the ACL-SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46, Washington, DC.
Mohammad, Saif and Graeme Hirst. Distributional measures as proxies for semantic relatedness. Submitted.
Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.
Newman, John and Sally Rice. 2004. Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics*, 15(3):351–396.
Nicolas, Tim. 1995. Semantics of idiom modification. In Everaert et al., editors, *Idioms: Structural and Psychological Perspectives*. LEA, Mahwah, NJ, pages 233–252.
Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.
Odijk, Jan. 2004. A proposed standard for the lexical representations of idioms. In *Proceedings of Euralex'04*, pages 153–164, Lorient.
Ogden, Charles Kay. 1968. *Basic English, International Second Language*. Harcourt, Brace, and World, New York.
Patrick, Jon and Jeremy Fletcher. 2005. Classifying verb-particle constructions by verb arguments. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 200–209, Colcheter.
Pauwels, Paul. 2000. *Put, Set, Lay and Place: A Cognitive Linguistic Approach to Verbal Meaning*. LINCOM EUROPA, Munich.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

R 2004. *Notes on R: A Programming Environment for Data Analysis and Graphics*. Available at `www.r-project.org`.

Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, (11):95–130.

Riehemann, Susanne. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford University.

Ritz, Julia and Ulrich Heid. 2006. Extraction tools for collocations and their morphosyntactic specificities. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1925–30, Genoa.

Rohde, Douglas L. T. 2004. *TGrep2 User Manual*. Available at `http://tedlab.mit.edu/~dr/Tgrep2`.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)*, pages 1–15, Mexico City.

Schenk, André. 1995. The syntactic behavior of idioms. In Everaert et al., editors, *Idioms: Structural and Psychological Perspectives*. LEA, Mahwah, NJ, chapter 10, pages 253–271.

Seaton, Maggie and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition, New York.

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

Tanabe, Harumi. 1999. Composite predicates and phrasal verbs in *The Paston Letters*. In L. J. Brinton and M. Akimoto. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. John Benjamins Publishing Company, Amsterdam, pages 97–132.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language*, 19:497–512.

Van de Cruys, Tim and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the ACL'07 Workshop on a Broader Perspective on Multiword Expressions*, pages 25–32, Prague.

Venkatapathy, Sriram and Aravid Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 899–906, Vancouver.

Villada Moirón, Begoña and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL'06 Workshop on Multiword Expressions in a Multilingual Context*, pages 33–40, Trento.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of multiword expressions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona.

Wermter, Joachim and Udo Hahn. 2005. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 843–850, Vancouver.

Widdows, Dominic and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of ACL'05 Workshop on Deep Lexical Acquisition*, pages 48–56, Ann Arbor, MI.

Wilcoxon, Frank. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024

Downloaded from http://direct.mit.edu/coli/article-pdf/35/1/61/1798560/coli.08-010-r1-07-048.pdf by guest on 14 November 2024