Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources

Yulia Tsvetkov*
Carnegie Mellon University

Shuly Wintner**
University of Haifa

We propose a framework for using multiple sources of linguistic information in the task of identifying multiword expressions in natural language texts. We define various linguistically motivated classification features and introduce novel ways for computing them. We then manually define interrelationships among the features, and express them in a Bayesian network. The result is a powerful classifier that can identify multiword expressions of various types and multiple syntactic constructions in text corpora. Our methodology is unsupervised and language-independent; it requires relatively few language resources and is thus suitable for a large number of languages. We report results on English, French, and Hebrew, and demonstrate a significant improvement in identification accuracy, compared with less sophisticated baselines.

1. Introduction

Multiword expressions (MWEs) are lexical items that consist of multiple orthographic words (*ad hoc, New York, look up*). MWEs constitute a significant portion of the lexicon of any natural language (Jackendoff 1997; Erman and Warren 2000; Sag et al. 2002). They are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words and other are phrases; some occur in one rigid pattern (and a fixed order), and others permit various syntactic transformations. The most characteristic property of MWEs is their semantic opacity, although the compositionality of MWEs is gradual, and ranges from fully compositional to completely idiomatic (Bannard, Baldwin, and Lascarides 2003).

* Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213-3891. E-mail: ytsvetko@cs.cmu.edu.
** Department of Computer Science, University of Haifa Mount Carmel, 31905 Haifa, Israel. E-mail: shuly@cs.haifa.ac.il.

Submission received: 6 January 2013; revised submission received: 13 June 2013; accepted for publication: 16 August 2013.

doi:10.1162/COLI_a_00177

Because of their prevalence and irregularity, MWEs must be stored in lexicons of natural language processing (NLP) applications. Awareness of MWEs was proven beneficial for a variety of applications, including information retrieval (Doucet and Ahonen-Myka 2004), building ontologies (Venkatsubramanyan and Perez-Carballo 2004), text alignment (Venkatapathy and Joshi 2006), and machine translation (Baldwin and Tanaka 2004; Uchiyama, Baldwin, and Ishizaki 2005; Carpuat and Diab 2010).

We propose a novel architecture for identifying MWEs, of various types and syntactic categories, in monolingual corpora. Unlike much existing work, which focuses on a particular syntactic construction, our approach addresses MWEs of various types by zooming in on the general idiosyncratic properties of MWEs rather than on specific properties of each subclass thereof. Addressing multiple types of MWEs has its limitations: The task is less well-defined, one cannot rely on specific properties of a particular construction, and the type of the MWE is not extracted along with the candidate expression. Nevertheless, there are clear benefits to such an approach. Certain applications can benefit from a large, albeit untyped, mixed bag of MWEs; machine translation is an obvious candidate (Lambert and Banchs 2005; Ren et al. 2009; Bouamor, Semmar, and Zweigenbaum 2012). Another use, which motivates our current work, is the construction of computational lexicons. Clearly, manual supervision is required before MWE candidates are added to a high-precision lexicon, but our approach provides the lexicographer with a large-scale set of potential candidates.

We focus on bigrams only in this work, that is, on MWEs consisting of two consecutive tokens. Many of the features we design, as well as the general architecture, can in principle be extended to longer MWEs, but we do not address longer (and, in particular, the harder case of non-contiguous) MWEs here. The architecture uses **Bayesian networks** (Pearl 1985) to express multiple interdependent linguistically motivated features.

First, we automatically generate a small (training) set of MWE and non-MWE bigrams (positive and negative instances, respectively) from a small parallel corpus. We then define a set of linguistically motivated features that embody observed characteristics of MWEs. We augment these by features that reflect collocation measures. Finally, we define dependencies among these features, expressed in the structure of a Bayesian network model, which we then use for classification. A Bayesian network (BN) is a directed graph whose nodes express the features used for classification and whose edges define causal relationships among these features. In this architecture, learning does not result in a black box, expressed solely as feature weights. Rather, the structure of the BN allows us to study the impact of different MWE features on the classification. The result is a new method for identifying MWEs of various types in text corpora. It combines statistics with an array of linguistically motivated features, organized in an architecture that reflects interdependencies among the features.

The contribution of this work is manifold.[1] First, we use existing approaches to MWE extraction to automatically generate training material. Specifically, we use our earlier work (Tsvetkov and Wintner 2012) to extract a set of positive and negative MWE candidates from a small parallel corpus, and use them for training a BN that can then extract a new set of MWEs from a potentially much larger *mono*lingual corpus. As

1 This article is a thoroughly revised and extended version of Tsvetkov and Wintner (2011). Whereas the methodology of that paper required minor supervision, we now present a completely unsupervised approach. We added several linguistically motivated features to the classification task. We demonstrate results on two new languages, English and French, to emphasize the generality of the method. Additional extensions include a more complete literature survey and, because new languages are added, different, more reliable data sets for evaluating our results.

a result, our method is completely unsupervised (more precisely, it does not require manual annotation; we do need several language resources, see Section 3.2).

Second, we propose several linguistically motivated features that can be computed from data and that are demonstrably productive for improving the accuracy of MWE identification. These features focus on the expression of linguistic idiosyncrasies of various types, a phenomenon typical of MWEs. Some of these features are commonplace, but others are new, or are implemented in novel ways. In particular, we account for the morphological idiosyncrasy of MWEs using a histogram of the number of inflected forms, in a technique that draws from image processing. We also use frequency histograms to model the semantic contexts of MWEs.

Finally, the methodology we advocate is not language-specific; given relatively few language resources, it can be easily adapted to new languages. We demonstrate the generality of our methodology by applying it to three languages: English, French, and Hebrew. Our evaluation shows that the use of linguistically motivated features results in a reduction of between one quarter and one third of the errors compared with a collocation baseline; organizing the knowledge in a Bayesian network reduces the error rate by an additional 3–9%.

After discussing related work in the next section (borrowing from Tsvetkov and Wintner [2012]), we motivate in Section 3 the methodology we propose, and list the resources needed for implementing it. Section 4 discusses the linguistically motivated features and their implementation; the organization of the Bayesian network is described in Section 5. We explain how we generate training materials in Section 6. Section 7 provides a thorough evaluation of the results. We conclude with suggestions for future research.

2. Related Work

Early approaches to MWE identification concentrated on their collocational behavior (Church and Hanks 1990). One of the first approaches was implemented as Xtract (Smadja 1993): Here, word pairs that occur with high frequency within a context of five words in a corpus are first collected, and are then ranked and filtered according to contextual considerations, including the parts of speech of their neighbors. Pecina (2008) compares 55 different association measures in ranking German Adj-N and PP-Verb collocation candidates. He shows that combining different collocation measures using standard statistical classification methods improves over using a single collocation measure. Other results (Chang, Danielsson, and Teubert 2002; Villavicencio et al. 2007) suggest that some collocation measures (especially point-wise mutual information and log-likelihood) are superior to others for identifying MWEs.

Co-occurrence measures alone are probably not enough to identify MWEs, and their linguistic properties should be exploited as well (Piao et al. 2005). Hybrid methods that combine word statistics with linguistic information exploit morphological, syntactic, and semantic idiosyncrasies to extract idiomatic MWEs.

Cook, Fazly, and Stevenson (2007), for example, use prior knowledge about the overall syntactic behavior of an idiomatic expression to determine whether an instance of the expression is used literally or idiomatically. They assume that in most cases, idiomatic usages of an expression tend to occur in a small number of canonical forms for that idiom; in contrast, the literal usages of an expression are less syntactically restricted, and are expressed in a greater variety of patterns, involving inflected forms of the constituents.

Ramisch et al. (2008) evaluate a number of association measures on the task of identifying English verb-particle constructions and German adjective-noun pairs. They show that adding linguistic information (mostly POS and POS-sequence patterns) to the association measure yields a significant improvement in performance over using pure frequency.

Several works address the **lexical fixedness** or **syntactic fixedness** of (certain types of) MWEs in order to extract them from texts. An expression is considered lexically fixed if replacing any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or literal expression. Syntactically fixed expressions prohibit (or restrict) syntactic variation. For example, Van de Cruys and Villada Moirón (2007) use lexical fixedness to extract Dutch verb-noun idiomatic combinations (VNICs). Bannard (2007) uses syntactic fixedness to identify English VNICs. Another work uses both the syntactic and the lexical fixedness of VNICs in order to distinguish them from non-idiomatic ones, and eventually to extract them from corpora (Fazly and Stevenson 2006). Recently, Green et al. (2011) use parsing, and in particular Tree Substitution Grammars, for identifying MWEs in French.

Semantic properties of MWEs can be used to distinguish between compositional and non-compositional (idiomatic) expressions. Katz and Giesbrecht (2006) and Baldwin et al. (2003) use Latent Semantic Analysis (LSA) for this purpose. They show that compositional MWEs appear in contexts more similar to their constituents than non-compositional MWEs. For example, the co-occurrence measured by LSA between the expression *kick the bucket* and the word *die* is much higher than co-occurrence of this expression and its component words. The disadvantage of this methodology is that to distinguish between idiomatic and non-idiomatic usages of the MWE it relies on the MWE's known idiomatic meaning, and this information is usually not available. In addition, this approach fails when only idiomatic or only literal usages of the MWE are overwhelmingly frequent.

Although these approaches are in line with ours, they require lexical semantic resources (e.g., a database that determines semantic similarity among words) and syntactic resources (parsers) that are unavailable for many languages. Our approach only requires morphological processing and a bilingual dictionary, which are more readily available for several languages. Note also that these approaches target a specific syntactic construction, whereas ours is appropriate for various types of MWEs.

Several properties of Hebrew MWEs are described by Al-Haj (2010); Al-Haj and Wintner (2010) use them in order to construct a support vector machine (SVM) classifier that can distinguish between MWE and non-MWE *noun-noun constructions* in Hebrew. The features of the SVM reflect several morphological and morphosyntactic properties of such constructions. The resulting classifier performs much better than a naive baseline, reducing the error rate by over one third. We rely on some of these insights, as we implement more of the linguistic properties of MWEs. Again, our methodology is not limited to a particular construction: Indeed, we demonstrate that our general methodology, trained on automatically generated, general training data, performs almost as well as the noun-noun-specific approach of Al-Haj and Wintner (2010) on the very same data set (Section 7).

Recently, Tsvetkov and Wintner (2010b, 2012) introduced a general methodology for extracting MWEs from bilingual corpora, and applied it to Hebrew. The results were a highly accurate set of Hebrew MWEs, of various types, along with their English translations. A major limitation of this work is that it can only be used to identify MWEs in the bilingual corpus, and is thus limited in its scope. We use this methodology to extract both positive and negative instances for our training set in the current work;

but we extrapolate the results much further by extending the method to *mono*lingual corpora, which are typically much larger than bilingual ones.

Probabilistic graphical models are widely used in statistical machine learning in general, and natural language processing in particular (Smith 2011). Bayesian networks are an instance of such models, and have been used for classification in several natural language applications. For example, BNs have been used for POS tagging of unknown words (Peshkin, Pfeffer, and Savova 2003), dependency parsing (Savova and Peshkin 2005), and document classification (Lam, Low, and Ho 1997; Calado et al. 2003; Denoyer and Gallinari 2004). Very recently, Ramisch et al. (2010) used BN for Portuguese MWE identification. The features used for classification were of two kinds: (1) various collocation measures; (2) bigrams aligned together by an automatic word aligner applied to a parallel (Portuguese–English) corpus. A BN was used to combine the predictions of the various features on the test set, but the structure of the network is not described. The combined classifier resulted in a much higher accuracy than any of the two methods alone. However, the use of BN is not central to this work, and its structure does not reflect any insights or intuitions on the structure of the problem domain or on interdependencies among features.

We, too, acknowledge the importance of combining different sources of knowledge in the hard task of MWE identification. In particular, we also believe that collocation measures are highly important for this task, but cannot completely solve the problem: Linguistically motivated features are crucial in order to improve the accuracy of the classifier. In this work we focus on various properties of different types of MWEs, and define general features that may accurately apply to some, but not necessarily all, of them. An architecture of Bayesian networks is optimal for this task: It enables us to define weighted dependencies among features, such that certain features are more significant for identifying some class of MWEs, whereas others are more prominent in identifying other classes (although we never predefine these classes). As we show herein, this architecture results in significant improvements over a more naive combination of features.

3. Methodology

3.1 Motivation

The task we address is identification of MWEs, of various types and syntactic constructions, in monolingual corpora. These include proper names, noun phrases, verb-particle pairs, and so forth. We focus on bigrams (MWEs consisting of two consecutive tokens) in this work; the methodology, however, can be extended to longer n-grams. Several properties of MWEs make this task challenging: MWEs exhibit idiosyncrasies on a variety of levels, orthographic, morphological, syntactic, and of course semantic. Such a complex task calls for a combination of multiple approaches, and much research indeed suggests "hybrid" approaches to MWE identification (Duan et al. 2009; Hazelbeck and Saito 2010; Ramisch et al. 2010; Weller and Fritzinger 2010). We believe that Bayesian networks provide an optimal architecture for expressing various pieces of knowledge aimed at MWE identification, for the following reasons (noted, e.g., by Heckerman 1995):

- In contrast to many other classification methods, Bayesian networks can learn (and express) causal relationships between features. This facilitates better understanding of the problem domain.

- Bayesian networks can encode not only statistical data, but also prior domain knowledge and human intuitions, in the form of interdependencies among features (a possibility that we use here).

Furthermore, we try in this work to leverage the idiosyncrasy of MWEs and use it as a tool for identifying them.

Our definition of MWEs is operational: An expression is considered a MWE if it has to be stored in the lexicon of some NLP application; typically, this is because the expression exhibits some level of idiosyncratic behavior (semantic, syntactic, morphological, orthographic, etc.). In order to properly handle such expressions in downstream applications, the lexicon must store some specific information about the expression. This working definition motivates and drives our methodology: We leverage the idiosyncratic behavior of MWEs and define (Section 4) an array of features that capture and reflect this idiosyncrasy in order to extract MWEs from corpora.

3.2 Resources

Although our approach is in general not language-specific, applying it to any particular language requires several language resources which we specify in this section. In general, we require corpora (both monolingual and bilingual), morphological analyzers or stemmers, part-of-speech taggers, and bilingual dictionaries. No deeper processing is assumed (e.g., no parsers or lexical semantic resources are needed). The method we advocate is thus appropriate for *medium-density* languages (Varga et al. 2005).

To compute the features discussed in Section 4, we need large monolingual corpora. For English and French, we use the 10^9 corpora released for WMT-11 (Callison-Burch et al. 2011); the corpora were syntactically parsed using the Berkeley parser (Petrov and Klein 2007), but we only use the POS tags in this work. For Hebrew, we use a monolingual corpus (Itai and Wintner 2008), which we pre-process as in Tsvetkov and Wintner (2012): We use a morphological analyzer (Itai and Wintner 2008) to segment word forms (separating prefixes and suffixes) and induce POS tags. Summary statistics for each corpus are listed in Table 1.

For some features we need access to the lemma of word tokens. In Hebrew, the MILA morphological analyzer (Itai and Wintner 2008) provides the lemmas, but the parsed corpora we use in English and French do not. We therefore use the DELA dictionaries of English and French, available from LADL as part of the Unitex project (`http://www-igm.univ-mlv.fr/~unitex/`). The French dictionary lists 683,824 single-word entries corresponding to 102,073 lemmas, and 108,436 multiword entries corresponding to 83,604 MWEs. The English dictionary is smaller, with 296,606 single-word forms corresponding to 150,145 lemmas, and 132,990 multiword entries, corresponding

Table 1
Statistics of the monolingual corpora.

	English	French	Hebrew
Tokens	447,073,250	522,964,336	46,239,285
Types	2,421,181	2,416,269	188,572
Bigram tokens	429,550,149	505,441,224	45,858,152
Bigram types	22,929,768	21,428,007	5,698,581

Table 2
Statistics of the bilingual corpora.

	English–French		English–Hebrew	
Sentences	30,000	30,000	19,626	19,626
Tokens	834,707	895,632	271,787	280,508
Types	22,787	27,880	14,142	12,555
Bigram tokens	804,704	865,632	252,183	280,506
Bigram types	218,108	225,660	128,987	149,688

to 69,912 MWEs. If the corpus surface form is not listed in the dictionary, we use the surface form in lieu of its lemma. The multiword entries of the DELA dictionaries are only used for evaluation.

For some features we also need a bilingual dictionary. For English–Hebrew, we use a small dictionary consisting of 78,313 translation pairs. Some of the entries are collected manually, whereas others are produced automatically (Itai and Wintner 2008; Kirschenbaum and Wintner 2010). For English–French, because we are unable to obtain a good-quality dictionary, we use instead Giza++ (Och and Ney 2000) 1-1 word alignments computed automatically from the entire WMT-11 parallel corpus.

In order to prepare training material automatically (Section 6), we use small bilingual corpora. For English–French, we use a random sample of 30,000 parallel sentences from the WMT-11 corpus. For English-Hebrew, we use the parallel corpus of Tsvetkov and Wintner (2010a). Statistics of the parallel corpora are listed in Table 2.

For evaluation we need lists of MWEs, ideally augmented by lists of non-MWE bigrams. Such lists are notoriously hard to obtain. As a general method of evaluation, we run 10-fold cross-validation evaluation using the training materials (which we generate automatically). Additionally, we use three sets of MWEs for evaluation. First, we extract all the MWE entries from the English WordNet (Miller et al. 1990); we use the WordNet version that is distributed with NLTK (Bird, Klein, and Loper 2009). Second, we use the MWEs listed in the DELA dictionaries of English and French (see above). These sets only include positive examples, of course, so we only report recall results on them. For Hebrew, we use a small set that was used for evaluation in the past (Al-Haj and Wintner 2010; Tsvetkov and Wintner 2012). This is a small annotated corpus, **NN**, of Hebrew noun-noun constructions. The corpus consists of 413 high-frequency bigrams of the same syntactic construction; of those, 178 are tagged as MWEs (in this case, noun compounds) and 235 as non-MWEs. This corpus consolidates the annotation of three annotators: Only instances on which all three agreed were included. Because it includes both positive and negative instances, this corpus facilitates a robust evaluation of precision and recall.

4. Linguistically Motivated Features

We define several linguistically motivated features that are aimed at capturing some of the unique properties of MWEs. Although many idiosyncratic properties of MWEs have been previously studied, we introduce novel ways to express these properties as computable features that inform a classifier. Note that many of the features we describe in the following are completely language-independent; others are applicable to a wide range of languages, whereas few are specific to morphologically rich languages, and can

be exhibited in different ways in different languages. We provide examples in English, French, and Hebrew, drawn from the resources listed in Section 3.2. The methodology we advocate, however, is completely general.

A common theme for all these features is *idiosyncracy*: They are all aimed at locating some linguistic property on which MWEs may differ from non-MWEs. We begin by detailing these properties, along with the features that we define to reflect them. In all cases, the feature is applied to a **candidate MWE**, defined here as a bigram of tokens (all possible bigrams are potential candidates). The features are computed from the large monolingual corpora described in Section 3.2. In order for a feature to fire, at least five instances of the candidate MWE have to be present in the corpus.

Orthographic variation. Sometimes, MWEs are written with hyphens instead of inter-token spaces. Examples include Hebrew[2] *xd-cddi* (*one sided*) 'unilateral', English *elephant-bird*, and French *aide-soignant* (*help carer*) 'caregiver'. Of course, this feature is only relevant for languages that use the hyphen in this way.

We define a binary feature, HYPHEN, whose value is 1 iff the corpus includes instances of the candidate MWE in which the hyphen character connects the two tokens of the bigram.

Capitalization. MWEs are often named entities, and in languages such as English and French a large number of MWEs involve words whose first letter is capital. We therefore define a feature, CAPS, whose value is a binary vector with 1 in the i-th place iff the i-th word of the MWE candidate is capitalized.[3] For example, *the White House* will have the value $\langle 0, 1, 1 \rangle$. This feature is of course irrelevant for languages that do not use capitalization.

Fossil words. MWEs sometimes include constituents that have no usage outside the particular expression. Examples include Hebrew *ird lTmiwn* (*went-down to-treasury*) 'was lost', French *night club*, and English *hocus pocus*; as far as we know, this is a rather universal property.

We define a feature, FOSSIL, whose value is a binary vector with 1 in the i-th place iff the i-th word of the candidate only occurs in this particular bigram; the other words of the candidate expression can be morphological variants of each other, but must share the same lemma. For example, the value of FOSSIL for *hocus pocus* is $\langle 1, 1 \rangle$, whereas for French *night club* it is $\langle 1, 0 \rangle$. In order to filter out potential typos, candidates must occur at least five times in the corpus in order for this feature to fire.

Frozen form. MWE constituents sometimes occur in one fixed, frozen form, where the language's morphology licenses also other forms. For example, *spill the beans* does not license *spill the bean*, although *bean* is a valid form. Similarly, Hebrew *bit xwlim* (*house-of sick-people*) 'hospital' requires that the noun *xwlim* be in the plural; the variant *bit xwlh* (*house-of sick-person*) 'a sick person's house' only has the literal meaning. This feature is of use for languages that are not isolating.

2 To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmnsypcqršt*.

3 Here and in subsequent examples we do not assume that the length of an MWE is limited to 2. In the present work, however, the vector is of length exactly 2.

We define a feature, FROZEN, whose value is a binary vector with 1 in the i-th place iff the i-th word of the candidate never inflects in the context of this expression. For example, the value of FROZEN for *spill the beans* is $\langle 0, 1, 1 \rangle$, and for Hebrew *bit xwlim* (*house-of sick-people*) 'hospital' it is $\langle 0, 1 \rangle$.

Partial morphological inflection. In some cases, MWE constituents undergo a (strict but non-empty) subset of the full inflections that they would undergo in isolation. For example, the Hebrew *bit mšpT* (*house-of law*) 'court' occurs in the following inflected forms: *bit hmšpT* 'the court' (75%); *bit mšpT* 'a court' (15%); *bti hmšpT* 'the courts' (8%); and *bti mšpT* 'courts' (2%). Crucially, forms in which the second word, *mšpT* 'law,' is in the plural are altogether missing. Our assumption is that the inflection histograms of non-MWEs are more uniform than the histograms of MWEs, in which some inflections may be more frequent and others may be altogether missing. Of course, restrictions on the histogram may stem from the part of speech of the expression; such constraints are captured by dependencies in the BN structure.

We capture this property, which is again relevant for all non-isolating languages, with a technique that has been proven useful in the area of image processing (Jain 1989, Section 7.3). We compute a histogram of the distribution in the corpus of all the possible surface forms of each MWE candidate. Such histograms can compactly represent distributional information on morphological behavior, in the same way that histograms of the distribution of gray levels in a picture are used to represent the picture itself. For example, the histogram corresponding to *bit mšpT* (*house-of law*) 'court' would be

$$\langle (\textit{bit hmšpT}, 0.75), (\textit{bit mšpT}, 0.15), (\textit{bti hmšpT}, 0.08), (\textit{bti mšpT}, 0.02) \rangle$$

Because each MWE is idiosyncratic in its own way, we do not expect the histograms of MWEs to have some specific pattern, except non-uniformity. We therefore sort the columns of each histogram, thereby losing information pertaining to the specific inflections, and retaining only information about the idiosyncrasy of the histogram. For the example given, the obtained histogram is $\langle 75, 15, 8, 2 \rangle$. In contrast, the non-MWE *txwm mšpT* (*domain-of law*) 'domain of the law', which is syntactically identical, occurs in nine different inflected forms, and its sorted histogram is $\langle 59, 14, 7, 7, 5, 2, 2, 2, 2 \rangle$. The longer "tail" of the histogram is typical of compositional expressions.

Off-line, we compute the average histogram for positive and negative examples: The average histogram of MWEs is shorter and less uniform than the average histogram of non-MWEs. We define a binary feature, HIST, that determines whether the histogram of the candidate is closer, in terms of L_1 (Manhattan) distance, to the average histogram of positive or of negative examples.

In our corpora, the average histogram of English positive examples has exactly four elements: 93.62, 5.86, 0.45, and 0.05. This shows a clear tendency (93.62%) of English MWEs to occur in a single form only; and it also implies that *no* English MWE occurs in more than four variants. The English negative instances, in contrast, have a much longer histogram (12 elements); the first element is 85.83, much lower than the dominating element of the positive examples. In French, which is morphologically much richer, the number of elements in the average histogram of positive examples is 32 (the dominating elements are 90.8, 6.9, 1.1, 0.4), whereas the number of elements in the average histogram of negative examples is 92 (dominated by 75.6, 14.5, 3.9, 2.0).

Context. We hypothesize that MWEs tend to constrain their (semantic) context more strongly than non-MWEs. We expect words that occur immediately after MWEs to vary less freely than words that immediately follow other expressions. One motivation for this hypothesis is the observation that MWEs tend to be less polysemous than free combinations of words, thereby limiting the possible semantic context in which they can occur. This seems to us to be a universal property.

We define a feature, CONTEXT, as follows. We first compute a histogram of the frequencies of words following each candidate MWE. We trim the tail of the histogram by removing words whose frequency is lower than 0.1% (the expectation is that non-MWEs would have a much longer tail). Off-line, we compute the same histograms for positive and negative examples and average them as before. The value of CONTEXT is 1 iff the histogram of the candidate is closer (in terms of L_1 distance) to the positive average.

For example, the histogram of Hebrew *bit mšpT* 'court' includes 15 values, dominated by *bit mšpT yliwn* 'supreme court' (20%) and *bit mšpT mxwzi* 'district court' (13%), followed by contexts whose frequency ranges between 5% and 0.6%. In contrast, the non-MWE *txwm mšpT* 'domain-of law' has a much shorter histogram, namely (12, 11, 6): Over 70% of the words following this expression occur with frequency lower than 0.1% and are hence in the trimmed tail.

Syntactic diversity. MWEs can belong to various part of speech categories. We define as feature, POS, the category of the candidate, with values obtained by selecting frequent tuples of POS tags. For example, English *heart attack* is Noun-Noun, *dark blue* is Adj-Adj, *Al Capone* is PropN-PropN; French *chant funèbre* (*song funeral*) 'dirge' is Noun-Adj, *en bas* (*in low*) 'down' is Prep-Adj; Hebrew *rkbt hrim* (*train-of mountains*) 'roller-coaster' is Noun-Noun, and so forth.

Translational equivalents. Because MWEs are often idiomatic, they tend to be translated in a non-literal way, sometimes to a single word. We use a bilingual dictionary to generate word-by-word translations of candidate MWEs from Hebrew to English, and check the number of occurrences of the English literal translation in a large English corpus. For French–English, we check whether the literal translation occurs in the Giza++ (Och and Ney 2000) alignment results (we use *grow-diag-final-and* for symmetrization in this case, to improve the precision). Due to differences in word order between the two languages, we create two variants for each translation, corresponding to both possible orders. We expect non-MWEs to have some literal translational equivalent (possibly with frequency that correlates with their frequency in the source language), whereas for MWEs we expect no (or few) literal translations. For example, consider Hebrew *sprwt iph* (*literature pretty*) 'belles lettres'. Literal translation of the expression to English yields *literature pretty* and *pretty literature*; we expect these phrases to occur rarely in an English corpus. In contrast, the compositional *tmwnh iph* (*picture pretty*) 'pretty picture' is much more likely to occur literally in English.

We define a binary feature, TRANS, whose value is 1 iff some literal translation of the candidate occurs more than five times in the corpus. Although this feature is not language-specific, we assume that it should work best for pairs of rather distinct languages.

Collocation. As a baseline, statistical association measure, we use pointwise mutual information (PMI). We define a binary feature, PMI, with two values, *low* and *high*,

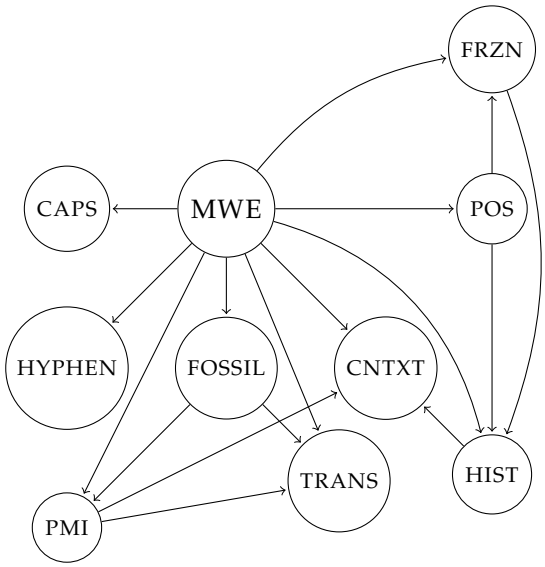

Figure 1
The Bayesian network for MWE identification.

reflecting an experimentally determined threshold. Clearly, other association measures (as well as combinations of more than one) could be used (Pecina 2005).

5. Feature Interdependencies Expressed as a Bayesian Network

A Bayesian network (Jensen and Nielsen 2007) is organized as a directed acyclic graph whose nodes are random variables and whose edges represent interdependencies among those variables. We use a particular view of BN, known as **causal** networks, in which directed edges lead to a variable from each of its direct *causes*.[4] This facilitates the expression of domain knowledge (and intuitions, beliefs, etc.) as structural properties of the network. We use the BN as a classification device: Training amounts to computing the joint probability distribution of the training set, and classification maximizes the posterior probability of the particular node (variable) being queried.

For MWE identification we define a BN whose nodes correspond to the features described in Section 4. In addition, we define a node, MWE, for the complete classification task. Over these nodes we impose the structure depicted graphically in Figure 1. This structure, which we motivate below, is *manually* defined: It reflects our understanding of the problem domain and is a result of our linguistic intuition. That said, it can of course be modified in various ways, and, in particular, new nodes can be easily added to reflect additional features.

All nodes depend on MWE, as all are affected by whether or not the candidate is an MWE. The POS of an expression influences its morphological inflection, hence the edges from POS to HIST and to FROZEN. For example, Hebrew noun-noun constructions allow their constituents to undergo the full inflectional paradigm, but when such a construction is a MWE, inflection is severely constrained (Al-Haj and Wintner 2010);

4 The direction of edges is from the target to the observable; this is compatible with the use of BNs in latent-variable generative models.

similarly, when one of the constituents of a MWE is a conjunction, the entire expression is very likely to be frozen, as in English *by and large* and *more or less*.

Fossil words clearly affect all statistical metrics, hence the edge from FOSSIL to PMI. They also affect the existence of literal translations, because if a word is not in the lexicon, it does not have a translation, hence the edge from FOSSIL to TRANS. Also, we assume that there is a correlation between the frequency (and PMI) of a candidate and whether or not a literal translation of the expression exists, hence the edge from PMI to TRANS. The edges from PMI and HIST to CONTEXT are justified by the correlation between the frequency and variability of an expression and the variability of the context in which it occurs.

Clearly, the process of determining the structure of the graph, and in particular the direction of some of the edges, is somewhat arbitrary. Having said that, it does give the designer of the system a clear and explicit way of expressing linguistically motivated intuitions about dependencies among features.

Once the structure of the network is established, the conditional probabilities of each dependency have to be determined. We compute the conditional probability tables from our training data (see Section 6) using Weka (Hall et al. 2009), and obtain values for $P(X \mid X_1, \ldots, X_k)$ for each variable X and all variables X_i, $1 \leq i \leq k$ (parents of X), such that the graph includes an edge from X_i to X. We then use the network for classification by maximizing $P(X_{mwe} \mid X_1, \ldots, X_k)$, where X_{mwe} corresponds to the node MWE, and $X_1, \ldots, X_k$ are the variables corresponding to all *other* nodes in the network. According to Bayes rule, we have

$$P(X_{mwe} \mid X_1, \ldots, X_k) \propto$$
$$P(X_1, \ldots, X_k \mid X_{mwe}) \times P(X_{mwe})$$

We define the prior, $P(X_{mwe})$, to be 0.41: This is the percentage of MWEs in WordNet 1.7 (Fellbaum 1998). This figure is of course rather arbitrary, but several studies indicate that the percentage of MWEs in the (mental) lexicon is approximately one half (Jackendoff 1997; Erman and Warren 2000; Sag et al. 2002). Post factum, we experimented with various other values for this parameter. We chose values between 0.3 and 0.55, in increments of 0.05, and computed the F-score of the system on the task of extracting English MWEs (see Section 7). As Table 3 shows, the differences are small (and *not* statistically significant), meaning that the accuracy of the system seems to be rather robust to the actual value of the prior. Given a small tuning set, it should be possible to optimize the choice of the prior more systematically.

The conditional probabilities $P(X_1, \ldots, X_k \mid X_{mwe})$ are determined by Weka from the conditional probability tables:

$$P(X_1, \ldots, X_k \mid X_{mwe}) = \Pi_{i=1}^{k} P(X_i \mid \mathbf{pa}_i)$$

where k is the number of nodes in the BN (other than X_{mwe}) and $\mathbf{pa}_i$ is the set of parents of X_i.

Table 3
F-score as a function of the value of the prior.

$P(X_{mwe})$	0.3	0.35	0.4	0.41	0.45	0.5	0.55
F-score	0.848	0.84	0.833	0.835	0.831	0.836	0.843

Table 4
Sizes of the training sets.

	MWE	non-MWE	Total
English	1,381	2,004	3,385
French	1,445	2,089	3,534
Hebrew	350	504	854

6. Automatic Generation of Training Data

For training we need samples of positive and negative instances of MWEs, each associated with a vector of the values of all features discussed in Section 4. We generate this training material automatically, using the small bilingual corpora described in Section 3.2. Each parallel corpus is first word-aligned with IBM Model 4 (Brown et al. 1993), implemented in Giza++ (Och and Ney 2003); we use *union* for symmetrization here, to improve the recall. Then, we apply the (completely unsupervised) algorithm of Tsvetkov and Wintner (2012), which extracts MWE candidates from the aligned corpus and re-ranks them using statistics computed from a large monolingual corpus.

The core idea behind this algorithm is that MWEs tend to be translated in non-literal ways; in a parallel corpus, words that are 1:1 aligned typically indicate literal translations and are hence unlikely constituents of MWEs. The algorithm hence focuses on *misalignments*: It trusts the quality of 1:1 alignments (which are further verified with a bilingual dictionary) and searches for MWEs exactly in the areas that word alignment *failed* to properly align, not relying on the alignment in these cases. Specifically, the algorithm views all words that are not included in 1:1 alignments as potential areas in which to search for MWEs, independently of how these words were aligned by the word-aligner. Then, it uses statistics computed from a large *monolingual* corpus to rank the MWE candidates; specifically, we use the PMI score of candidates based on counts from the monolingual corpora. Finally, the algorithm extracts maximally long sequences of words from the unaligned parallel phrases, in which each bigram has a PMI score above some threshold (determined experimentally). All bigrams in those sequences are considered MWEs. See Tsvetkov and Wintner (2012) for more details.

The set of MWEs that is determined in this way constitutes the positive examples in the training set. For negative examples, we use two sets of bigrams: Those that *are* 1:1 aligned and have high PMI; and those that are misaligned but have low PMI. To decide how many negative examples to generate, we rely on the ratio between MWE and non-MWE entries in WordNet, as mentioned above: $P(X_{mwe}) = 0.41$. We thus select from the negative set approximately 50% more negative examples than positive ones, such that the ratio between the sizes of the sets is 0.41 : 0.59. The sizes of the resulting training sets are listed in Table 4.

7. Results and Evaluation

We use the training data described in Section 6 for training and evaluation: We perform 10-fold cross validation experiments, reporting accuracy and (balanced) F-score in three set-ups: One (SVM) in which we train an SVM classifier[5] with the features described

5 We use Weka SMO with the PolyKernel set-up; experimentation with several other kernels yielded worse results.

Table 5
10-fold cross validation evaluation results.

	Hebrew		French		English	
	Accuracy (%)	F-score	Accuracy (%)	F-score	Accuracy (%)	F-score
PMI	66.98	0.67	70.88	0.762	74.15	0.737
BN-auto	71.19	0.71	77.45	0.775	82.16	0.822
SVM	74.59	0.75	78.38	0.736	82.95	0.828
BN	**76.82**	**0.77**	**79.04**	**0.778**	**83.52**	**0.835**

in Section 4; one (BN-auto) in which we train a Bayesian network with these features, but let Weka determine its structure (using the K2 algorithm); and one (BN) in which we train a Bayesian network whose structure reflects manually crafted linguistically motivated knowledge, as depicted in Figure 1. The results are listed in Table 5; they are compared with a PMI baseline, obtained by defining a Bayesian network with only two nodes, MWE and PMI.

The linguistically motivated features defined in Section 4 are clearly helpful in the classification task: The accuracy of an SVM, informed by these features, is close to 75% for Hebrew, over 78% for French, and as high as 83% for English, reducing the error rate of the PMI baseline by 23% (Hebrew) to 34% (English). The contribution of the BN is also highly significant, reducing 3–9% more errors (with respect to the errors made by the SVM classifier).[6] In total, the best method, **BN**, reduces the error rate of the PMI-based classifier by one third. Interestingly, a BN whose structure does not reflect prior knowledge, but is rather learned automatically, performs worse than these two methods (but still much better than relying on PMI alone).[7] It is the combination of linguistically motivated features with feature interdependencies reflecting domain knowledge that contribute to the best performance.

We did not investigate the contribution of each of the features to the classification task. However, we did analyze the weights assigned by the SVM classifier to specific features. Unsurprisingly, the most distinctive feature is PMI. Among the POS features, the strongest feature is *VB_NNS*, an indication of a negative instance. Capitalization is also unsurprisingly a very strong feature. We leave a more systematic analysis of the contribution of each feature to future work.

To further assess the quality of the results, we performed a human evaluation on the English data set. We first produced the results in the BN set-up, and then sorted both the (predicted) positive and the (predicted) negative instances by their PMI. We randomly picked 100 instances of both lists, at the same positions in the ranked lists, to constitute an evaluation set. We asked three English-speaking annotators to determine whether the 200 expressions were indeed MWEs. The annotation guidelines are given in Appendix A. Comparing the three annotators' labels, we found out that they agreed on 141 of the 200 (70.5%). This should probably be taken as an upper bound for the task.

6 The improvement of both BN and SVM over the baseline is highly significant statistically (sign test, $p < 0.01$ in all three cases); the improvement of BN over SVM is significant for English ($p < 0.01$) but not for French.

7 We are not sure why this is the case. One possible explanation is that our training set contains noisy examples, and as the BN-auto classifier learns the dependencies from noisy data, it performs worse than the SVM classifier. Another possible explanation is that it attempts to learn more dependencies, thereby increasing the parameter space of the model.

We then computed the majority label and compared it with our predicted label. Exactly 142 of the predicted labels were annotated as correct; that's an accuracy of 71%. Of the 141 instances that the three annotators agreed on, our results predict the correct label for 112 instances (79.4%). We take these figures as a strong indication of the accuracy of the results.

As an additional evaluation measure, we use the sets of bigrams in the English WordNet, and the bigram MWEs in the DELA dictionaries of English and French (Section 3.2). Because we only have positive instances in these evaluation sets, we can only report recall. We therefore use the Bayesian network classifier to extract MWEs from the large monolingual corpora discussed in Section 3.2. For each evaluation set (WordNet, DELA English, and DELA French), we divide the number of bigrams in the set that are classified as MWEs by the size of the intersection of the evaluation set with the monolingual corpus. In other words, we exclude from the evaluation those MWEs in the evaluation set that never occur in our corpora. The results are listed in Table 6.

As examples of correctly identified MWEs, consider English *advisory board, air cargo, adoption agency, air ticket, crude oil*, and so on, and French *accord international* 'international agreement', *acte final* 'final act', *banque centrale* 'central bank', *ce soir* 'tonight', and so forth, all taken from the DELA dictionaries. The relatively low recall of our method on these dictionaries is to a large extent due to a very liberal definition of MWEs that the dictionaries use. Many entries that are listed as MWEs are actually highly compositional, and hence our method fails to identify them. DELA entries that are *not* identified by our classifier include examples such as English *abnormal behavior, absolute necessity, academic research*, and so on. The French DELA dictionary is especially extensive, with examples such as *action sociale, action antitumorale, action associative, action caritative, action collective, action commerciale, action communautaire*, and many more, all listed as MWEs. Our system only recognizes the first of these.

The WordNet results are obviously much better. Correctly identified MWEs include *ad hoc, outer space, web site, inter alia, road map*, and so forth. WordNet MWEs that our system failed to identify include *has been, as well, in this, a few, set up*, and so on. A more involved error analysis is required in order to propose potential directions for improvement on this set.

As a further demonstration of the utility of our approach, we evaluate the algorithm on the set **NN** of Hebrew noun-noun constructions described in Section 3.2. We train a Bayesian network on the training set described in Section 6 and use it to classify the set **NN**. We compare the results of this classifier with a PMI baseline, and also with the classification results reported by Al-Haj and Wintner (2010); the latter reflects 10-fold cross-validation evaluation using the entire set, so it may be considered an upper bound for any classifier that uses a *general* training corpus.

The results are depicted in Table 7. They clearly demonstrate that the linguistically motivated features we define provide a significant improvement in classification accuracy over the baseline PMI measure. Note that our F-score, 0.77, is very close to the

Table 6
Evaluation results: WordNet and DELA dictionaries.

	True positives	Evaluation set size	Recall (%)
WordNet	25,549	42,403	60
DELA English	11,955	26,460	45
DELA French	886	4,798	18

Table 7
Evaluation results: noun-noun constructions.

	Accuracy	Precision	Recall	F-score
PMI	71.43%	0.71	0.71	0.71
BN	**77.00%**	**0.77**	**0.77**	**0.77**
AW	80.77%	0.77	0.81	0.79

AW = results from Al-Haj and Wintner (2010)

best result of 0.79 obtained by Al-Haj and Wintner (2010) as the average of 10-fold cross validation runs, using *only* high-frequency noun-noun constructions for training. We interpret this result as a further proof of the robustness of our architecture.

Finally, we conduct an analysis of the quality of extracted (Hebrew) MWEs. We used the trained BN to classify the entire set of bigrams present in the (Hebrew side of the) Hebrew–English parallel corpus described in Section 3.2. Of the more than 140,000 candidates, only 4,000 are classified as MWEs. We sort this list of potential MWEs by the probability assigned by the BN to the positive value of the variable X_{mwe}. The resulting sorted list is dominated by high-PMI bigrams, especially proper names, all of which are indeed MWEs. The first non-MWE (false positive) occurs in the 50th place on the list; it is *crpt niqwla* 'France Nicolas', which is obviously a sub-sequence of the larger MWE, *neia crpt niqwla srqwzi* 'French president Nicolas Sarkozy'. Similar sub-sequences are also present, but only five are in the top 100. Such false positives can be reduced when longer MWEs are extracted, as it can be assumed that a sub-sequence of a longer MWE does not have to be identified. Other false positives in the top 100 include some highly frequent expressions, but over 85 of the top 100 are clearly MWEs.

Although more careful evaluation is required in order to estimate the rate of true positives in this list, we trust that the vast majority of the positive results are indeed MWEs.

8. Conclusions and Future Work

We presented a novel architecture for identifying MWEs in text corpora. The main insights we emphasize are sophisticated computational encoding of linguistic knowledge that focuses on the idiosyncratic behavior of such expressions. This is reflected in two ways in our work: by defining computable features that reflect different facets of irregularities; and by framing the features as part of a larger Bayesian network that accounts for interdependencies among them. We also introduce a method for automatically generating a training set for this task, which renders the classification entirely unsupervised. The result is a classifier that can identify MWEs of several types and constructions. Evaluation on three languages (English, French, and Hebrew) shows significant improvement in the accuracy of the classifier compared with less sophisticated baselines.

The modular architecture of Bayesian networks facilitates easy exploration with more features. We are currently investigating the contribution of various other sources of information to the classification task. For example, Hebrew lacks large-scale lexical semantic resources. However, it is possible to literally translate an MWE candidate to English and rely on the English WordNet for generating synonyms of the literal translation. Such "literal synonyms" can then be back-translated to Hebrew. The assumption is

that if a back-translated expression has a low PMI, the original candidate is very likely not a MWE. Although such a feature may contribute little on its own, incorporating it in a well-structured BN may improve performance. Another feature that can easily be implemented in this way is whether the POS of MWE constituents is retained when the expression is translated to another language; we hypothesize that this is much more likely when the expression is compositional.

Appendix A. Annotation Guidelines

These are the instructions given to the annotators.

> The task is to annotate each line as either a multi-word expression, in which case mark 1 in the first field; or not, in which case the value is 0. It's a hard task, but you are requested to be decisive. Please do not change the file in any other way.
>
> The main criterion for determining whether an expression is a MWE is whether it has to be stored in a computational lexicon. Typically, expressions are stored in lexicons if they exhibit idiosyncratic (irregular) behavior. This could be due to:
>
> - non-compositional meaning. For example, 'green light' is an MWE because it is not a light. 'kill time' is not a violent action. A good indication of non-compositional meaning is limited reference. For example, if someone gives you a green light, you can't then refer to it as 'the light I was given'.
> - non-substitutability of elements. For example, 'breast cancer' is an MWE because while 'breast' and 'chest' can often be substituted, 'breast cancer' and 'chest cancer' cannot.
> - fossil words, i.e., words that only occur in the context of the expression. For example, 'mutatis mutandis'.
> - nominalization. If the expression can occur as a single word, or with a connecting hyphen, it is a strong indication that it is an MWE. For example, 'road map' can be written 'roadmap'.
> - irregular syntactic and/or morphological behavior. For example, 'look up' is an MWE because while ordinarily you can convert 'I walked up the alley' to 'Up the alley I walked', you can't convert 'I looked up that word in a dictionary' to 'Up that word I looked'.
> - proper names. All proper names are by definition MWEs. This includes people ('Barack Obama'), places ('Tel Aviv'), organizations ('United Nations'), etc.
>
> But really, the best criterion is: if I hadn't known this expression, would I be able to use it properly simply by knowing its two constituents? Would I understand its meaning, be able to inflect it properly, construct syntactic constructions, and in general use it in the right context in the right way?

Acknowledgments

This research was supported by The Israel Science Foundation (grants 137/06 and 1269/07). We are grateful to Gennadi Lembersky for his continuous help, and to the three anonymous Computational Linguistics reviewers for very constructive comments that greatly improved this article. All remaining errors are of course our own.

References

Al-Haj, Hassan. 2010. Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition. Master's thesis, University of Haifa.

Al-Haj, Hassan and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic

idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 10–18, Beijing.
Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, pages 89–96, Sapporo.
Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona.
Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the 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 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.
Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.
Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 674–679, Istanbul.
Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
Calado, Pável, Marco Cristo, Edleno Silva De Moura, Nivio Ziviani, Berthier A. Ribeiro-Neto, and Marcos André Gonçalves. 2003. Combining link-based and content-based methods for web document classification. In *Proceedings of CIKM-03, 12th ACM International Conference on Information and Knowledge Management*, pages 394–401, New Orleans, LA.
Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh.
Carpuat, Marine and Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, CA.
Chang, Baobao, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, pages 1–5, Morristown, NJ.
Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
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 Workshop on a Broader Perspective on Multiword Expressions (MWE 2007)*, pages 41–48, Prague.
Denoyer, Ludovic and Patrick Gallinari. 2004. Bayesian network model for semi-structured document classification. *Information Processing and Management*, 40(5):807–827.
Doucet, Antoine and Helana Ahonen-Myka. 2004. Non-contiguous word sequences for information retrieval. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, Barcelona.
Duan, Jianyong, Mei Zhang, Lijing Tong, and Feng Guo. 2009. A hybrid approach to improve bilingual multiword expression extraction. In Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*. Springer, Berlin and Heidelberg, pages 541–547.
Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text*, 20(1):29–62.
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)*, pages 337–344, Trento.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press, Cambridge, MA.

Green, Spence, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Hazelbeck, Gregory and Hiroaki Saito. 2010. A hybrid approach for functional expression identification in a Japanese reading assistant. In *Proceedings of the 2010 Workshop on Multiword Expressions: From Theory to Applications*, pages 81–84, Beijing.

Heckerman, David. 1995. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA.

Itai, Alon and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Jain, Anil K. 1989. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ.

Jensen, Finn V. and Thomas D. Nielsen. 2007. *Bayesian Networks and Decision Graphs*. Springer, 2nd edition.

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney.

Kirschenbaum, Amit and Shuly Wintner. 2010. A general method for creating a bilingual transliteration dictionary. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 273–276, Valletta.

Lam, Wai, Kon F. Low, and Chao Y. Ho. 1997. Using a Bayesian network induction approach for text categorization. In *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence*, pages 745–750, Nagoya.

Lambert, Patrik and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of the MT Summit X*, pages 396–403, Phuket.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. *International Journal of Lexicography*, 3(4):235–312.

Och, Franz Josef and Hermann Ney. 2000. Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pearl, Judea. 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society*, pages 329–334, University of California, Irvine, CA.

Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, MI.

Pecina, Pavel. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech.

Peshkin, Leonid, Avi Pfeffer, and Virginia Savova. 2003. Bayesian nets in syntactic categorization of novel words. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–Short Papers - Volume 2*, NAACL '03, pages 79–81, Edmonton.

Petrov, Slav and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, pages 404–411, Rochester, NY.

Piao, Scott Songlin, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language*, 19(4):378–397.

Ramisch, Carlos, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria Finatto. 2010. A hybrid approach for multiword expression identification.

In Thiago Pardo, António Branco, Aldebaro Klautau, Renata Vieira, and Vera de Lima, editors, *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*. Springer, Berlin and Heidelberg, pages 65–74.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Alline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pages 50–53, Marrakech.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54, Singapore.

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

Savova, Virginia and Leonid Peshkin. 2005. Dependency parsing with dynamic Bayesian network. In *Proceedings of the 20th National Conference on Artificial Intelligence—Volume 3*, pages 1,112–1,117, Pittsburgh, PA.

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

Smith, Noah A. 2011. *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Tsvetkov, Yulia and Shuly Wintner. 2010a. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 3,389–3,392, Valletta.

Tsvetkov, Yulia and Shuly Wintner. 2010b. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1256–1264, Beijing.

Tsvetkov, Yulia and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 836–845, Edinburgh.

Tsvetkov, Yulia and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(4):549–573.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech & Language*, 19(4):497–512.

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

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596, Borovet.

Venkatapathy, Sriram and Aravind Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27, Sydney.

Venkatsubramanyan, Shailaja and Jose Perez-Carballo. 2004. Multiword expression filtering for building knowledge. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 40–47, Barcelona.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1,034–1,043, Prague.

Weller, Marion and Fabienne Fritzinger. 2010. A hybrid approach for the identification of multiword expressions. In *Proceedings of the SLTC 2010 Workshop on Compounds and Multiword Expressions*, pages 1–2, Linköping.