Universität des Saarlandes
Philosophische Fakultät II
Allgemeine Linguistik
Computerlinguistik

Master's Thesis

Identification of Idiomatic Expressions Using Parallel Corpora

Anna Mündelein
October 29, 2008

First supervisor: Dr. Caroline Sporleder
Second supervisor: Prof. Dr. Manfred Pinkal

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Saarbrücken, 29. Oktober 2008

Anna Mündelein

Contents

1 Introduction

In the last years, the detection and classification of multiword expressions (MWEs) has received much attention. MWEs pose problems to almost all natural language processing applications. In machine translation, treating MWEs as single words leads to incorrect (and often laughable) translations. In part-of-speech tagging, tagging MWEs can result in implausible part-of-speech sequences; *by and large*, for example, could get the tags "preposition conjunction adjective". A dialog system that does not consider MWEs will perform badly when trying to understand the user input. Other applications that benefit from the correct treatment of MWEs are parsing, terminology extraction, bilingual/multilingual MWE alignment, automatic interpretation and generation of language, paraphrasing, and summarization.

Villada Moirón and Tiedemann (2006) present a type-based method for ranking Dutch MWEs of the form PP verb according to their idiomaticity. They use parallel corpora and automatic word-alignment (Dutch-German, Dutch-English, and Dutch-Spanish) and base their method on the hypothesis that compositional expressions are translated consistently while non-compositional expressions get more diverse alignments. The diversity of an alignment is calculated as translational entropy. In addition, the proportion of default alignments for an MWE is calculated.

To test the feasibility of the approach by Villada Moirón and Tiedemann (2006) in a wider context, we re-implemented it for German and extended it in several ways. As well as MWEs of the form PP verb, we analyzed those of the form NP verb. In addition to the languages Villada Moirón and Tiedemann (2006) used, we explored the German-Finnish alignment because we hypothesized that the approach might work better for very distant languages that are less likely to contain the same MWEs. When calculating the scores,

we experimented with combining the scores from the different language pairs. We also tried to make the approach token-based such that it classifies the single occurrences of an MWE as literal or idiomatic. This was done by extracting all occurrences of the MWE in question from the corpus, clustering them and the MWE into two clusters, and calculating translational entropy for both clusters. The cluster with the higher translational entropy score was considered the idiomatic cluster and the other one the literal cluster. If the idiomatic cluster contained the MWE in question, it was classified as idiomatic; otherwise it was classified as literal.

The remainder of the thesis is structured as follows: In chapter 2, MWEs and their subclasses idioms and support verb constructions are defined. In chapter 3, we will shortly present the statistical measures of association that are often used for MWEs. Chapter 4 gives an overview about related approaches. Chapter 5 presents our type-based and chapter 6 our token based method. In chapter 7, there will be a short conclusion.

2 Multiword Expressions

Multiword expressions (MWEs) can be defined as "two or more words that together form a single unit of meaning" (Fazly and Stevenson, 2007) or as "cohesive lexemes that cross word boundaries" (Baldwin *et al.*, 2003). Some examples for MWEs are *post office, play with fire, make a decision*, and *push on*.

2.1 Decomposability

According to Riehemann (2001) and Sag *et al.* (2002), decomposability is "a description of the degree to which the semantics of an MWE can be ascribed to those of its parts". Baldwin *et al.* (2003) distinguish three classes of MWEs: non-decomposable, idiosyncratically decomposable, and simple decomposable MWEs.

Non-decomposable MWEs are MWEs for which no decompositional analysis is possible, for example *kick the bucket, shoot the breeze*, or *hot dog*. The only syntactic variation they allow for is verbal inflection (*kicked the bucket, kicks the bucket*).

Idiosyncratically decomposable MWEs are decomposable, but coerce their parts into taking semantics unavailable outside the MWE. In *spill the beans*, for example, *spill* is coerced into *reveal* and *the beans* into *a secret*, meanings they usually do not have, only in the context of the MWE. Other examples for this kind of MWE are *let the cat out of the bag* and *radar footprint*. Idiosyncratically decomposable MWEs allow for a certain degree of syntactic variation, for example passive voice (*the cat was let out of the bag*).

Simple decomposable MWEs like *kindle excitement* or *traffic light* decompose into simplex senses. Even though they show a high degree of syntactic variation, there is a difference to productive word combinations: Simple decomposable MWEs tend to block compositional alternates with the expected semantics, as you can see in example 1. This test also works for the other kinds of MWEs.

(1) motor car, *engine car, *motor automobile

2.2 Linguistic Modeling

There are two main approaches to the linguistic modeling of MWEs: the words-with-spaces approach and the compositional approach.

The **words-with-spaces** approach models MWEs as single lexical entries. This works for non-decomposable MWEs that are completely syntactically fixed, for example *by and large*. Non-decomposable MWEs that undergo verbal inflection as well as idiosyncratically decomposable and simple decomposable MWEs pose problems to this approach. For MWEs like these, we need a **compositional** approach that treats MWEs by general and compositional methods of linguistic analysis.

2.3 Idioms

Idioms are a subset of MWEs; they are either non-decomposable or idiosyncratically decomposable.

Higi-Wydler (1989) gives a three-part definition of idioms:

1. Idioms are fixed word combinations and are reproduced as a whole.

2. Idioms consist of at least two words, but are smaller than a sentence.

3. The meaning of the idiom as a whole does not (fully) correlate with the sum of the meanings of its components.

An important property of idioms is their figurativeness. According to Daniels (1963), they are "situation images" (*Situationsbilder*) that once referred to a special action. This can be transparent to the speaker or not: In *unter die Lupe nehmen* ("to take under the magnifying glass" — to examine carefully), for example, it is clear to which image the idiom refers, but in idioms like *Kohldampf haben* ("to have cabbage steam" — to be starving) it is not.

Idioms take different positions within a sentence and can be distinguished with respect to their syntactic functions. Higi-Wydler (1989) enumerates five idiom classes: prepositional and conjunctional, adjectival, nominal, adverbial, and verbal idioms.

Prepositional idioms consist of nouns that are headed by prepositions. An example for such an idiom is *im Lauf(e)* (in the course of).

(2) Jedes Los hat eine Nummer. Acht dieser Nummern werden *im Laufe* des Abends unter notarieller Aufsicht als Haupttreffer ermittelt.
[St. Galler Tagblatt, May 2, 1997]

An example for a **conjunctional idiom**, an idiom that functions as a conjunction, is *gesetzt den Fall (dass)* (be it (that)).

(3) Aber *gesetzt den Fall*, er will nicht bleiben, würden Sie dann als Nachfolger zur Verfügung stehen?
[Tiroler Tageszeitung, December 30, 1996]

Adjectival idioms are word combinations with adjectives or participles that occur in predicative (4) as well as in attributive (5) positions. Examples

are *fix und fertig* (pooped) or *bis an die Zähne bewaffnet* (armed to the teeth).

(4) Als Zuschauer ist man nach diesen 130 Minuten fast so *fix und fertig* wie die Protagonistin Isabelle Huppert, die mit ihrem Schauspiel alles gegeben hat.
[St. Galler Tagblatt, November 9, 2001]

(5) Das Schattenkabinett um Rugova, das sich auf einen breiten Konsens in der albanischen Regierung stützt, übt praktisch die Regierungsgewalt aus, der Belgrad nur *bis an die Zähne bewaffnete* Polizeibataillone entgegenzusetzen hat.
[Salzburger Nachrichten, September 28, 1994]

Nominal idioms are idioms that are headed by a noun. Examples are *ein hohes/großes Tier* (a big shot), *das kalte Buffet*, (the cold buffet), or *der kleine Mann* (ordinary Joe). Usually, they are used as subjects or objects, but they also occur as genitive attributes (6), predicatives (7), and adverbials (8). If a word combination only occurs with *sein* (to be) or *haben* (to have), for example *der Hahn im Korb sein* (to be cock of the roost), Higi-Wydler (1989) does not classify it as a nominal, but as a verbal idiom.

(6) Als Kraut ist's uns ein etwas gröberer (und den ganzen Sommer durch erhältlicher) Spinatersatz, die Stiele gehen als Spargel *des kleinen Mannes* durch.
[St. Galler Tagblatt, May 16, 1997]

(7) Die Schlenderer sind der *Sand im Getriebe* des städtischen Alltags.
[St. Galler Tagblatt, March 12, 1998]

(8) Für rund 850 Millionen Schilling soll ein Fünf-Sterne-Haus mit allem *Drum und Dran* aus dem Boden wachsen.
[Tiroler Tageszeitung, April 29, 2000]

Auf der Stelle (right away), *letzten Endes* (eventually), *in der Regel* (as a rule), *unter vier Augen* (in confidence), and *zwischen Tür und Angel* (in passing) belong to the class of **adverbial idioms**. For distinguishing between adverbial and verbal idioms, the same rule as for nominal idioms is applied: If an idiom can occur without a verb, it is an adverbial idiom. If it only occurs with a special verb (or a number of synonymous verbs), it is a verbal idiom.

(9) Auch die alte Frau, die schließlich das stählerne Tor in der Saint-Georges-Street 7 einen Spalt öffnet, riskiert *zwischen Tür und Angel* nur wenige Worte.
[Frankfurter Rundschau, May 16, 1997]

Some adverbial idioms consist of whole subordinate clauses: *ohne mit der Wimper zu zucken* (without batting an eye), *dass die Fetzen fliegen* (that sparks fly), *wenn es hochkommt* (at best).

Verbal idioms constitute the biggest idiom class. One can distinguish between four subclasses:

1. Verb noun combinations are the most common verbal idioms. They can occur either with or without preposition: *den Hof machen* (to court), *in die Hand nehmen* (to take into one's own hands). This is the group of idioms on which we will concentrate in this work.

2. There are also verbal idioms that consist of a verb and an adjective or adverb: *satt haben* (to be fed up), *nicht schlau werden* (not to be

able to make head or tail), *hoch und heilig versprechen* (to promise solemnly).

3. Some idioms consist of two verbs, sometimes in connection with some other element, for example a noun: *dran glauben müssen* (to bite the dust), *sich keine grauen Haare wachsen lassen* (not to trouble one's head), *etwas läuten hören* (to hear something).

4. A special class is constituted by idioms that take the function of a sentence, but bring a subordinate clause with them: *Es ist an der Zeit, dass* (it is time), *Es wird höchste Zeit (dass)* (it is high time).

2.4 Support Verb Constructions

Kuhn (1994) defines **support verb constructions (SVCs)**, another subclass of MWEs, as follows: "A support verb construction is the combination of a support verb and a nominal component (possibly introduced by a preposition), where the overall meaning comes from the nominal component. The support verb contributes only general semantic information, like tense, aspect, aktionsart and causation."

According to Bahr (1977), SVCs can be divided in the five groups shown in Table 1. One can see that the distinction between SVCs and idioms (groups A and A') on the one hand and between SVCs and free word combinations (groups E and E') on the other hand is not entirely clear.

Group A comprises SVCs with properties of an idiom: They show the same syntactic-morphologic restrictions. The difference to group A' is that the expressions in group A are no (full) "situation images" anymore; their nominal part can be seen as something abstract, in contrast to *Braten, Leim, Bockshorn, Leine*, etc.

Group A can be divided into two subgroups: The nouns in example 10 only occur in connection with certain verbs. The nouns in example 11 can occur as independent lexemes in free contexts, yet together with certain verbs, they do not realize the meanings they usually have. The meanings can only be inferred from the expressions as a whole.

(10) in Betracht ziehen
in Betracht kommen
außer Betracht lassen
zum Vorschein kommen
Obacht geben
Obacht haben
in Abrede stellen
Vorschub leisten

(11) in Angriff nehmen
ins Gebet nehmen
in Kauf nehmen
zum Zuge kommen
am Zuge sein
in Rechnung stellen
Rechnung tragen
das Nachsehen haben

Group B consists of phraseological SVCs with strict syntactic-morphologic restrictions. In contrast to the groups A and A', the nominal part has a clear meaning on its own — the meaning it also has in other contexts. [1]

[1]It is not clear why Bahr (1977) puts *Reißaus nehmen* in group B, as *Reißaus* does not seem to have a clear meaning on its own.

A’	ins Bockshorn jagen	idioms	
	den Braten riechen		
	Leine ziehen		
	auf den Leim gehen		
A	in Kauf nehmen		SVCs
	in Angriff nehmen		
	Vorschub leisten		
	Rechnung tragen		
B	Folge leisten		
	den Vorzug geben		
	zum Stehen bringen/kommen		
	Reißaus nehmen		
C	Einfluss nehmen		
	Rache nehmen		
	in Not geraten		
	in den Ruin treiben		
D	eine Wendung nehmen		
	eine Regelung treffen		
	eine Frage stellen		
	einen Vorschlag machen		
E	eine Überprüfung vornehmen	free word combinations	
	eine Überprüfung erfolgt		
	Reue empfinden		
E’	eine Überprüfung verweigern		
	eine Überprüfung beginnt		
	seine Reue demonstrieren		
	einen Vorschlag ausführen		

Table 1: Distinction between idioms, SVCs, and free word combinations

Group C are phraseological SVCs with loosened syntactic-morphologic restrictions. For these expressions, it is possible to add attributes (12).

(12) *großen* Einfluss nehmen
beträchtlichen Schaden nehmen
schmerzlichen Abschied nehmen
seinen *gewohnten* Lauf nehmen

In group D, one finds phraseological SVCs with substantially loosened syntactic-morphologic restrictions. In addition to adding attributes, it is possible to add a relative clause to the noun (13).

(13) der Antrag, *den* ich stelle
die Rede, *die* ich halte

Group E, finally, comprises non-phraseological SVCs without syntactic-morphologic restrictions.

3 Statistical Measures

In this chapter, we will shortly present the lexical association measures that have shown to be important for the detection and classification of MWEs: χ^2, log-likelihood, and mutual information. These measures are often taken as a comparison when evaluating more elaborated measures. We will base our description on Manning and Schütze (1999).

3.1 χ^2

The χ^2 test compares the observed frequencies of a group of events with the frequencies expected for independence of these events. If the difference between observed and expected frequencies is large, the null hypothesis of independence can be rejected.

χ^2 is defined as follows:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{14}$$

i and j range over the rows and columns of a frequency table, O_{ij} is the observed value for cell (i, j), and E_{ij} is the expected value. An example for a frequency table can be seen in Table 2.

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 *(new companies)*	4667 (e.g. *old companies*)
$w_2 \neq companies$	15820 (e.g. *new machines*)	14287181 (e.g. *old machines*)

Table 2: A 2-by-2 table showing the dependence of occurrences of *new* and *companies* in a corpus

3.2 Log-likelihood

Likelihood ratios are more interpretable and more appropriate for sparse data than the X^2 test. They give us a number that tells us how much more likely one hypothesis is than the other.

For collocation discovery, there are the following two hypotheses for the occurrence frequency of a bigram w^1w^2:

- $H_1 : P(w^2|w^1) = p = P(w^2|\neg w^1)$
- $H_2 : P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

H_1 is a formalization of independence and H_2 a formalization of dependence. $p = \frac{c_2}{N}$, $p_1 = \frac{c_{12}}{c_1}$, and $p_2 = \frac{c_2 - c_{12}}{N - c_1}$, where N is the total number of words, and c_1, c_2, and c_{12} are the numbers of occurrences of w^1, w^2, and w^1w^2, respectively, in the corpus.

The log of the likelihood ratio λ is then:

$$\begin{aligned} log\lambda &= log\frac{L(H_1)}{L(H_2)} \\ &= logL(c_{12}, c_1, p) + logL(c_2 - c_{12}, N - c_1, p) \\ &- logL(c_{12}, c_1, p_1) - logL(c_2 - c_{12}, N - c_1, p_2) \end{aligned} \tag{15}$$

where $L(k, n, x) = x^k(1 - x)^{n-k}$.

3.3 Mutual Information

Mutual information (MI) is a symmetric, non-negative measure of the common information in two variables. It is 0 when the two variables are independent, and for two dependent variables, it grows not only with the degree of dependence, but also according to the entropy of the variables.

MI is defined as follows:

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{16}$$

Pointwise mutual information (PMI) is defined as:

$$I(x,y) = log \frac{p(x,y)}{p(x)p(y)} \tag{17}$$

4 Related Work

The work about MWEs can be roughly divided into type-based and token-based methods.

Type-based methods classify expression types rather than expression tokens, that is, an expression is classified irrespective of the context in which it occurs. A type-based method would, for example, say that *play with fire* is an idiom and *play with marbles* is not.

Token-based methods, in contrast, consider the concrete instances of an abstract type. This is useful because many idioms can be used idiomatically as well as literally. The analysis of 60 idioms by Cook *et al.* (2007) showed that close to half of them have a clear literal meaning, and of the expressions with a literal meaning, on average 40% of their usages are literal. A token-based method would classify the occurrence of *play with fire* in example 18 as idiomatic and the occurrence in example 19 as literal.

(18) And even the low oil price scenario could lead to political instability in oil producing countries that may rebound against the West and oil consuming countries. "You are *playing with fire*," he warned.
[BBC News, March 14, 2003]

(19) Fire officers said the incident illustrated the dangers of children *playing with fire*.
[BBC News, April 21, 2006]

We will present some type-based and some token-based approaches. In the end of this chapter, we will talk in more detail about the type-based approach by Villada Moirón and Tiedemann (2006) that this work is based upon.

4.1 Type-based Approaches

The type-based approaches can be grouped into five classes, according to the linguistic properties or methods they use. Some exploit the lexical fixedness of MWEs, some their syntactic fixedness, and some the distributional similarity between an expression and its components. One approach combines several measures into a set of features, and there are also approaches that use evidence from another language to detect MWEs.

Using Lexical Fixedness

One of the most influential works on the non-compositionality of MWEs is that of Lin (1999). His method is based on the hypothesis that when an expression is non-compositional, the MI value between its components differs significantly from the MI values of expressions that result from substituting one of the words in the expression with a similar word.

Lin (1999) automatically builds a thesaurus and a collocation database. For each collocation, he substitutes each of the component words with a word with similar meaning, taken from the thesaurus. The MI value is calculated by taking a collocation as the combination of three events: the type of the dependency relation, the head lexical item, and the modifier. A phrase α is taken to be non-compositional if and only if there exists no phrase β such that: (a) β can be produced by substitution of the components of α, and (b) there is an overlap between the 95% confidence interval of the MI values of α and β. These judgments are compared to a dictionary of idioms.

Lin (1999) reports quite low scores of 15.7% for precision and 13.7% for recall.

Like Lin (1999), Wermter and Hahn (2005) explore the lexical fixedness of

non-compositional MWEs. They extract bigrams, trigrams, and quadrams from a biomedical text corpus. Then they measure their *limited paradigmatic modifiability*, the probability with which one or more slots of an expression cannot be filled by other tokens.

Their method outperforms the standard approaches for distinguishing terms from non-terms in the biomedical literature (t-test and C-value).

Using Syntactic Fixedness

Fazly and Stevenson (2007) and Bannard (2007) try to detect MWEs by analyzing their syntactic fixedness.

Fazly and Stevenson (2007) distinguish three kinds of MWEs: idioms, light verb constructions (which are the same as SVCs), and so-called *abstract combinations* like *make a living*. They define and measure some linguistic properties: *Institutionalization* is defined as "the process through which a combination of words becomes recognized and accepted as a semantic unit involving some degree of semantic idiosyncrasy". This can be applied to all three kinds of MWEs and is measured as their PMI. *Lexicosyntactic fixedness* refers to "some degree of lexical and syntactic restrictiveness in a semantically idiosyncratic expression", so it only applies to idioms and light verb constructions. *Non-compositionality* is defined as "the situation where the meaning of a word combination deviates from the meaning emerging from a word-by-word interpretation of it". It can mostly be applied to idioms, but also to light verb constructions, and it is measured as the cosine similarity between the context vector of the expression and the context vectors of its parts.

The highest performance, an accuracy of 58% (the baseline is 25%), is

achieved when combining all the properties.

Bannard (2007) focuses on verb noun combinations and especially on phrases that allow some syntactic variation, but do not allow other kinds. There are three important kinds of non-morphological variation that such phrases can undergo: modification of the noun phrase, passivization of the verb phrase, and variation, addition, or dropping of a determiner.

Bannard (2007) estimates the extent to which the probability of variation for a combination deviates from the variation that would be expected based on the variation that is observed for its component words. This is measured as conditional PMI. The total information about the syntactic variation for a word pair is calculated as the sum of the information gained about all verb-relevant syntactic variations and the information gained about all object relevant variations.

Evaluation is done using a gold standard set of dictionary MWEs, and it shows that the method proposed can identify this set with greater accuracy than the lexical association measures t-score, MI, log-likelihood, and χ^2. Also, the method extracts very different items than these measures.

Using Distributional Similarity Between an Expression and its Components

Other approaches examine the distributional similarity between an expression and its components, for example McCarthy *et al.* (2003), Baldwin *et al.* (2003), and Bannard *et al.* (2003).

McCarthy *et al.* (2003) parse the British National Corpus (BNC) and let humans rank the detected 111 phrasal verbs according to their composition-

ality. They automatically build a thesaurus with the 500 nearest neighbors for the set of phrasal verbs and also for their simplex verbs. Then they use various measures, for example overlap, that compare the nearest neighbors of a phrasal verb to the nearest neighbors of its corresponding simplex verb.

The evaluation shows that there is a significant relationship between the human compositionality judgement and some of the measures — a stronger relationship than with the statistics that have previously been used for filtering candidate multiwords (χ^2, log-likelihood, and PMI).

Baldwin *et al.* (2003) want to differentiate non-decomposable, idiosyncratically decomposable, and simple decomposable MWEs. Their idea is to compare the distribution of different compound terms with their constituent words to see if this indicates similarity of meaning. Latent semantic analysis (LSA) is used to model meaning: Each word is represented as a vector of its context words. The similarity between a word and its constituents is measured as the cosine of the angle between the corresponding vectors.

The evaluation is done using WordNet: For simple decomposable MWEs, the constituents are expected to be hypernyms or synonyms of the MWE in the WordNet hierarchy. The results suggest that there is a moderate correlation between the LSA similarities and occurrences of hyponymy in WordNet.

According to Bannard *et al.* (2003), MWE compositionality can be defined as an entailment relationship between the whole and its parts. The challenge is that MWEs do not fall cleanly into binary classes of compositional and non-compositional expressions, but populate a continuum between the two extremes.

They build a classifier with four different methods: 1. a re-implementation

of Lin's method (Lin, 1999), 2. a knowledge-free approach which does not use a thesaurus, 3. a method based on the corpus-based similarity between the original expression and word-substituted derivative expressions, and 4. a method based on the intuition that identifying the degree of semantic similarity between a verb-particle construction and its component verb and/or particle will indicate whether that component part contributes independent semantics.

The results show that on all tasks at least one of the methods offers an improvement in precision over the baseline.

Combining Several Measures

The approach of Venkatapathy and Joshi (2005) combines lexical fixedness, distributional similarity, and collocation-based measures into a set of features. They use a support vector machine (SVM) based function to integrate the features and rank verb noun collocations according to their compositionality. These ranks are then compared to human annotation.

The evaluation shows that the method outperforms traditional measures. Especially the features "similarity of the collocation to the verb-form of the object", "least MI difference with similar collocations", and "distributed frequency of object using the verb information" contribute greatly to measuring the compositionality.

Using Evidence from Another Language

The approach of Melamed (1997a) relies on the idea that non-compositional compounds (NCCs) are usually not translated literally to other languages. That means, one way to discover NCCs is to induce and analyze a translation model between the languages. This translation model is constructed auto-

matically from bitexts in two phases: First, the algorithm finds and counts the most likely links between word tokens in the two halves of the bitext, and then it estimates translation probabilities by dividing the link counts by the total number of links.

There is a simple test to see if an NCC is valid: Two translation models are induced, one that involves the candidate NCC and one that does not. If the value of the objective function is higher in the model with the NCC, then the NCC is valid.

Villada Moirón and Tiedemann (2006) also use evidence from another language to detect MWEs. Since this is the approach our work is based upon, it will be presented in more detail in subsection 4.3.

4.2 Token-based Approaches

There are much fewer token-based than type-based approaches. We will present three of them: one that uses clustering, one that uses the local linguistic context of a token, and one that uses syntactic forms.

Using Clustering

Birke and Sarkar (2006) develop Trope Finder (TroFi), a nearly unsupervised clustering method for separating literal and non-literal usages of verbs. *Literal* is defined here as "anything that falls within accepted selectional restrictions". The problem of non-literal language recognition is reduced to one of word sense disambiguation, and an existing similarity-based word sense disambiguation method is adapted to this task.

The TroFi algorithm requires a *target set* and two *seed sets*: a *literal feedback set* and a *non-literal feedback set*. The target set is a set of sen-

tences that contain the verbs to be classified. The seed sets are collected from WordNet and other sources and contain feature lists consisting of the stemmed nouns and verbs in a sentence, with target/seed words and frequent words removed.

The algorithm is based on the principle of attraction: Similarities are calculated between the sentences containing the target word and the feedback sets. A target set sentence is attracted to the feedback set containing the sentence to which it shows the highest similarity. Two sentences are similar if they contain similar words, and two words are similar if they are contained in similar sentences. The target set sentences are clustered into literal and non-literal, depending on which feature set they are attracted to.

The feedback sets are very noisy because they are collected automatically. For this reason, Birke and Sarkar (2006) apply learners and a voting system to clean the feedback sets. In addition, so-called *SuperTags* are used to add internal structure information to the feedback sets.

For evaluation, the target set sentences are annotated by two humans as literal or non-literal. The baseline is a simple attraction algorithm: A target set sentence is attracted to the feedback set with which it has the most words in common. On average, 7.6% improvement in F-score over the baseline are achieved.

Using Local Linguistic Context

Katz and Giesbrecht (2006) propose a method for using the local linguistic context of an expression to determine if it is used literally or idiomatically. Like Birke and Sarkar (2006), they see an analogy to the two tasks of word sense disambiguation: sense discrimination and selection. Sense discrimination consists of determining for a given expression if it has a non-

compositional interpretation in addition to its compositional interpretation (type-based classification), and selection consists of determining in a given context if a given expression is used compositionally or non-compositionally (token-based classification). The meaning of an expression is modeled in terms of the words it co-occurs with, using LSA.

Katz and Giesbrecht (2006) conduct two experiments:

In the first experiment, 67 occurrences of *ins Wasser fallen* ("to fall into the water" — to be canceled) are annotated by two humans as literal or idiomatic. LSA meaning vectors for both the literal and the idiomatic usages are calculated. As expected, the vectors are almost orthogonal, with a cosine of the angle between them of 0.02. Then a 10-fold cross-validation study is conducted. The literal and idiomatic vectors for *ins Wasser fallen* are calculated on the basis of the training data, and a simple nearest neighbor classification is done for each member of the test set on the basis of the meaning vectors computed from its local context. The resulting average accuracy of 72% outperforms the simple maximum-likelihood baseline of 58%.

In the second experiment, Katz and Giesbrecht (2006) want to classify 81 potential MWEs according to whether or not they have an idiomatic meaning. The estimated compositional meaning vector of an MWE is computed by summing up the meaning vectors of its parts. This vector is then compared to the actual vector of the MWE as a whole. If the similarity between the two vectors is low, the MWE is classified as having a non-compositional usage. For a similarity threshold of 0.2, an F-measure of 0.48 is achieved.

Using Syntactic Forms

The approach of Cook *et al.* (2007) relies on the idea that prior knowledge about the overall syntactic behavior of an idiomatic expression (type-based

knowledge) can be used to determine if the expression is used literally or idiomatically in a certain context (token-based knowledge). Based on this hypothesis, unsupervised methods for token classification are developed, concentrating on verb noun combinations like *make a face*.

The *canonical forms* of an idiom are defined as its most preferred syntactic patterns. Each idiom has only one or a small number of canonical forms.

Idioms tend to be fixed in regard to the syntactic configurations in which they occur. Literal usages of an expression are less syntactically restricted and show a greater variety of patterns. That means, if one finds an MWE that appears in one of its canonical forms, it is likely to be of idiomatic usage, and if one finds an MWE that appears in a form different from its canonical forms, it is likely to be of literal usage.

For finding the canonical forms of an idiom, Cook *et al.* (2007) use an unsupervised statistical method: The canonical forms are the forms whose frequency is much higher than the average frequency of all of its forms.

They define a so-called *CForm method* and two *Diff methods*: The CForm method classifies a token as idiomatic if it occurs in one of the canonical forms for that expression, and as literal otherwise. The two Diff methods combine local context information with knowledge of the canonical form. They take a distributional approach to meaning: The meaning of an expression is approximated by the words with which it co-occurs. Co-occurrence vectors are built to represent this meaning. The assumption is that the idiomatic and literal usages of an expression correspond to two coarse-grained senses of the expression, being used with different sets of words and thus resulting in different co-occurrence vectors.

The first Diff method builds two co-occurrence vectors, one for all uses of the expression in its canonical form(s) and one for all uses of the expression

in its non-canonical forms. The first vector represents the idiomatic and the second vector the literal meaning.

The second Diff method also builds a co-occurrence vector for all uses of the expression in its canonical form(s), representing the idiomatic meaning. As the meaning of a literal expression can be composed of the meanings of its parts, the literal meaning vector is built by summing and then normalizing the co-occurrence vectors for the expression's component words.

Both Diff methods classify a token as idiomatic if its vector is more similar (using the cosine similarity measure) to the idiomatic meaning vector than to the literal meaning vector, and as literal otherwise.

The results for the approach are almost as good as for the supervised approach of Katz and Giesbrecht (2006) and always better than the baseline (always predicting the idiomatic label). The CForm method shows the best result with 72.4% accuracy. This means that the canonical forms are very important for distinguishing between idiomatic and literal usages of an expression.

4.3 Identification of Idiomatic Expressions Using Parallel Corpora

Villada Moirón and Tiedemann (2006) present a type-based method to create rankings for Dutch MWEs according to their idiomaticity. Like Melamed (1997a), they use evidence from other languages for their approach. In chapter 5, we will show how we re-implemented their method for German MWEs.

Idea

According to Villada Moirón and Tiedemann (2006), deciding if an expression is idiomatic or literal can be done in two ways: by measuring how predictable

the meaning of the expression is and by assessing the link between the meaning of the expression as a whole and the cumulative literal meanings of its components. Parallel corpora and automatic word-alignment can be used for this purpose. It is assumed that compositional MWEs are translated consistently and that an expression is literal if its translation is the result of combining the translations of the single words. Also, Villada Moirón and Tiedemann (2006) claim that an automatic word-aligner will perform badly when trying to align idiomatic expressions word by word. They expect a large variety of links for each component word and hypothesize that these links are different from the default alignments found in the corpus.

Data and Resources

The Dutch part of Europarl (Koehn, 2005) and its translations to English, Spanish, and German are used. Europarl is a parallel corpus extracted from the proceedings of the European Parliament and contains about 40 million tokens in about 1.6 million sentences for Dutch.

The corpora are tokenized, sentence-aligned, and word-aligned with the program GIZA++ [2]. Word-alignments are produced for both translation directions, source to target (src2trg) and target to source (trg2src). Villada Moirón and Tiedemann (2006) also use an alignment that combines the two directional alignments, the so-called *refined alignment* (Och *et al.*, 1999).

The Dutch section of Europarl is parsed with the Alpino parser [3].

[2] http://www.fjoch.com/GIZA++.html

[3] http://www.let.rug.nl/~vannoord/alp/Alpino/source/

Triple	Links in English
aan	NO_LINK, to, of, in, for, from, on, into, at
licht	NO_LINK, light, revealed, exposed, highlight, shown, shed light, clarify
breng	NO_LINK, brought, bring, highlighted, has, is, makes

Table 3: Collection of translation links for the triple *an licht breng* (bring to light)

Extracting MWEs from the Corpus

In the parsed corpus, all sentences are chosen that contain at least one of the following main or support verbs: *doen, gaan, geven, hebben, komen, maken, nemen, brengen, houden, krijgen, stellen, zitten*. Out of these sentences, all verb PP tuples are collected, altogether 191,000 types and 413,000 tokens. To avoid data sparseness, the NP inside the PP is reduced to the head noun's lemma, and verbs are lemmatized too. Types that occur less than 50 times are ignored.

For each candidate triple, log-likelihood, salience, and head dependence between each PP and its selecting verbs are calculated. The resulting three rankings are combined to form one final ranking. From the final ranking, the top 200 triples are selected.

Collecting Alignments

For each word, all translation links to a target language, together with their frequencies, are collected. The resulting dictionary is called **link lexicon**.

Also, all translation links for each triple are collected. An example of this collection can be seen in Table 3. NO_LINK means that a word is not aligned to any other word.

Measures

As it is harder to align words in idiomatic expressions than other words, one can expect a larger variety of translation links. Entropy is a good measure for the unpredictability of an event. If there is a large variety of translation links, the entropy will be high.

Following Melamed (1997b), **translational entropy** is defined as in 20. s stands for the source word, t for the target word, and T_s for the set of target words that a source word is aligned to. $P(t|s)$ is the frequency of s and t together divided by the frequency of s.

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) log P(t|s) \tag{20}$$

The translational entropy of a triple is calculated as the average translational entropy of its components.

Villada Moirón and Tiedemann (2006) experiment with three variants of entropy: leaving out NO_LINKS, counting NO_LINKS as multiple types, and counting all NO_LINKS as one unique type.

The second measure used is the **proportion of default alignment (pda)**, which is calculated as in 21.

$$pda(S) = \frac{\sum_{s \in S} \sum_{d \in D_s} \textit{align-freq}(s, d)}{\sum_{s \in S} \sum_{t \in T_s} \textit{align-freq}(s, t)} \tag{21}$$

align-freq(s, t) is the alignment frequency of source word s to target word t in the context of the triple S. d stands for a word out of the set of default alignments. The default alignments are defined as the four most frequent alignment types extracted from the corresponding link lexicon.

The idea is that if an expression has a literal meaning, the translations of

its component words are similar to their default alignments. If an expression has an idiomatic meaning, the translations of its component words are very different from their default alignments.

Experiments and Results

Villada Moirón and Tiedemann (2006) experiment with the three word-alignment types (src2trg, trg2src, and refined) and the two measures (translational entropy and pda).

The 200 candidate MWEs are classified into idiomatic or literal by a human expert. Then the rankings resulting from the measures are compared to the expert's classification. The initial ranking of MWEs serves as the baseline.

Standard precision and recall are not applicable here because instead of defining an artificial cutoff for the ranked list, the ranking itself has to be evaluated. For this reason, Villada Moirón and Tiedemann (2006) calculate the so-called **uninterpolated average precision (uap)** (22) that aggregates precision points into one evaluation figure. At each point c where a true positive S_c in the ranked list is found, the precision $P(S_1 \dots S_c)$ is computed, and all precision points are then averaged. (Manning and Schütze, 1999)

$$uap = \frac{\sum_{S_c} P(S_1 \dots S_c)}{|S_c|} \tag{22}$$

In Table 4, one can see the comparison of the three different alignment types, using the entropy measure, for the Dutch-English alignment. The source to target alignment gives the best result.

Table 5 shows the comparison of the two different scoring metrics, translational entropy and pda, using the source to target alignment. All scores produce better rankings than the baseline. Pda performs slightly better than

Alignment	uap
src2trg	0.864
trg2src	0.785
refined	0.765
baseline	0.755

Table 4: Comparison of alignment types, using entropy, for NL-EN

Score	NL-EN	NL-ES	NL-DE
entropy			
- without NO_LINKS	0.864	0.892	0.907
- NO_LINKS = many	0.858	0.890	0.883
- NO_LINKS = one	0.859	0.890	0.911
pda	0.891	0.894	0.894
baseline	0.755	0.755	0.755

Table 5: Comparison of scoring metrics, using src2trg

translational entropy, except for the Dutch-German language pair.

Villada Moirón and Tiedemann (2006) try to further improve their method by lemmatizing the collected alignments and the link lexicon, but find out that lemmatization adds little to or even decreases the accuracy of the entropy and pda scores.

5 Type-based Approach

We took the same approach as Villada Moirón and Tiedemann (2006), with some exceptions that will be mentioned later. If not explicitly said otherwise, everything in this chapter is inspired by Villada Moirón and Tiedemann (2006).

5.1 Types of MWEs

In contrast to Villada Moirón and Tiedemann (2006), we analyzed not only MWEs of the form PP verb, for example *ans Herz legen* (to recommend warmly), but also MWES of the form NP verb, for example *Weichen stellen* (to set the course).

5.2 Languages

The Europarl corpus contains about 38 million tokens in about 1.5 million sentences for German. While Villada Moirón and Tiedemann (2006) analyzed MWEs in Dutch and used the translations to German, English, and Spanish for this purpose, we took German as our source language and used the translations to Dutch, English, and Spanish. In addition, we analyzed the German-Finnish word-alignment because we wanted to find out whether the results for a non-Indo-European language are better. It might be the case that in languages that are close to each other some MWEs are the same — for example *unter den Teppich kehren* and *to sweep under the carpet* —, so one will not find the diverse word-alignment that is necessary for the measures to work. In very distant languages we expect all MWEs to be different.

5.3 Automatic Alignment

First, the corpora for all language pairs had to be sentence-aligned. For this purpose, we needed a list of abbreviations for each language so that the sentence aligner knew where a sentence ended and where a period only signaled an abbreviation. For German and English, lists of abbreviations were delivered with the Europarl corpus. For Dutch, we found a list in Wiktionary [4], for Spanish in Ctspanish [5], and for Finnish in Kotus [6].

The corpora were aligned with the sentence aligner and tokenized with the tokenizer included in the Europarl tools. We merged the single corpora files into two big files, one for each language, while lowercasing the text and removing XML tags. We also removed empty lines, long sentences (longer than 60 words), and sentences with a high ratio between source and target sentence length (a ratio higher than five) because GIZA++ cannot cope with these.

After these operations, we used GIZA++ with standard settings to word-align the corpora. As Villada Moirón and Tiedemann (2006) got the best results for the source to target alignment, we only aligned from German to the other languages.

First, the source and target language texts were converted to a special format, *vocabulary files*, that GIZA++ needs as input. A *bitext file* and a *co-occurrence file* were created as well. Then GIZA++ was started with all the relevant files as parameters and produced an alignment file.

An excerpt of the German-English alignment file can be seen in Figure 1. The file is divided into triples: The first line gives the source and target sentence length and the alignment score (the probability that the alignment

[4] http://en.wiktionary.org/wiki/Category:Dutch_abbreviations
[5] http://www.ctspanish.com/words/abbreviations.htm
[6] http://www.kotus.fi/index.phtml?s=2149

was correct), the second line is the target, and the third line the source sentence. The numbers stand for the positions of the target words which the source words are aligned to. Figure 2 shows the alignment of the first sentence in a graphical way. There are mistakes in the alignment, even in such a small excerpt: *Wiederaufnahme* is aligned to both *resumption* and *of*, *der* is not aligned to *the*, etc.

```
# Sentence pair (1) source length 3 target length 4 alignment score : 9.92821e-05
resumption of the session
NULL ({ 3 }) wiederaufnahme ({ 1 2 }) der ({ }) sitzungsperiode ({ 4 })
# Sentence pair (2) source length 33 target length 40 alignment score : 2.08176e-73
i declare resumed the session of the european parliament adjourned on friday 17
december 1999 , and i would like once again to wish you a happy new year in the hope
that you enjoyed a pleasant festive period .
NULL ({ 4 6 7 16 17 23 26 30 36 }) ich ({ 1 18 19 20 }) erkläre ({ 2 3 10 }) die ({ })
am ({ }) freitag ({ 11 12 }) , ({ }) dem ({ }) 17. ({ 13 }) dezember ({ 14 })
unterbrochene ({ 15 }) sitzungsperiode ({ 5 }) des ({ }) europäischen ({ 8 })
parlaments ({ 9 }) für ({ }) wiederaufgenommen ({ }) , ({ }) wünsche ({ 24 }) ihnen
({ 25 }) nochmals ({ 21 22 }) alles ({ }) gute ({ }) zum ({ }) jahreswechsel ({ 27 28
29 31 }) und ({ }) hoffe ({ 32 33 }) , ({ }) daß ({ }) sie ({ 34 }) schöne ({ }) ferien
({ 35 37 38 39 }) hatten ({ }) . ({ 40 })
```

Figure 1: Excerpt of the alignment file

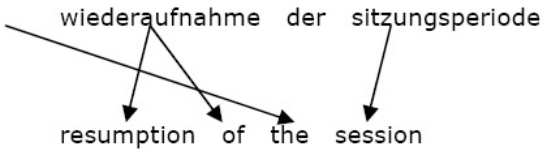

Figure 2: Alignment of the first sentence

5.4 Choosing Support Verbs

Since support verbs often occur with MWEs and Villada Moirón and Tiedemann (2006) chose support verbs for their analysis, we also wanted to use support verbs. For this purpose, we searched the Salsa Release Corpus (Burchardt *et al.*, 2006) for all verbs that had been annotated as support verbs.

The resulting list was extended by taking the support verbs out of the Tiger annotation scheme (Albert *et al.*, 2003). Then we counted the occurrences of all support verbs in the Europarl corpus, considering the different forms the verbs can have. The result can be seen in Table 6.

There were too many support verbs to consider all of them. We discarded *sein* and *haben* because they occur too often. The rest of the support verbs was divided into three groups, according to their frequency, in such a way that each group contained (almost) the same number of items. The partitioning of the support verbs can be seen in Table 6. From each group, we randomly chose five support verbs, so altogether we got fifteen verbs. These are: *geben, gehen, machen, stellen, kommen, fallen, ziehen, legen, treten, suchen, rufen, stoßen, treiben, üben*, and *stiften*.

5.5 Parsing

We wanted to extract verb NP and verb PP combinations from the Europarl corpus. To do this, we needed to know about the relations between the words — in other words, we needed a parser. As the Alpino parser, which Villada Moirón and Tiedemann (2006) used, only works for Dutch, we decided to use MaltParser [7], a data-driven dependency parser. For this parser, a parsing model has to be induced from a treebank in a given language and can then be used to parse this language.

We took the Tiger Corpus [8] as the training set. Both the training and the test set have to be in *CoNNL format* — an example of this format can be seen in Figure 3. The columns contain, in this order: a token counter, starting at one for each new sentence, the word form or punctuation symbol,

[7] http://w3.msi.vxu.se/~nivre/research/MaltParser.html

[8] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

Support verb	Frequency	Support verb	Frequency
sein	564585	bekommen	5403
haben	303291	treten	4596
geben	63121	erfahren	4543
gehen	38169	zukommen	4112
machen	37993	gewinnen	4070
stellen	37644	suchen	3724
kommen	30736	gelangen	3704
stehen	27450	vornehmen	3465
wissen	25545	besitzen	3430
lassen	23534	fassen	3418
finden	23386	erteilen	3099
führen	23335	lesen	2836
bringen	21557	erheben	2438
unternehmen	20745	lernen	2398
nehmen	20622	ausüben	2196
halten	20213	rufen	2126
sprechen	20160	geraten	2042
liegen	17471	schenken	1572
enthalten	13933	genießen	1423
bleiben	13555	stoßen	1348
erhalten	12898	wahrnehmen	1332
zeigen	12434	abgeben	1252
setzen	12364	empfinden	1087
treffen	12146	versetzen	1076
leisten	10439	melden	1029
heißen	10044	ausfallen	1019
schließen	10018	treiben	990
fallen	9028	üben	905
ergreifen	7723	einlegen	492
aufnehmen	7548	spannen	483
ziehen	7200	erregen	382
legen	6450	ablegen	268
darstellen	6349	stiften	229
befinden	5988	schöpfen	136
durchführen	5752	schweben	123
anstellen	5590	zuziehen	95
aussprechen	5406		

Table 6: List of support verbs with their frequencies in the Europarl corpus

the lemma or stem, a coarse-grained part-of-speech (POS) tag, a fine-grained POS tag, an unordered set of syntactic or morphological features, the head of the current token, the dependency relation to the head, the projective head of the current token, and the dependency relation to the projective head. Some columns might contain an underscore if the feature is not available.

After training the parser, we merged all German text files from the Europarl corpus into a single file, removing XML tags and empty lines. As the columns for the POS tags have to be present in the CoNNL format, we used TnT [9] to POS tag the corpus. In order to do this, we had to convert the file to a special format where each line contains only one word or punctuation symbol. After POS tagging, we converted the corpus to the CoNNL format, leaving out the columns for head and dependency relation.

Finally, the corpus could be parsed. This task would have taken around 300 days on one fast computer, that is why we split the corpus in thirteen parts and had them processed on several cluster nodes in parallel. In addition, we only considered those sentences that contain one of the chosen support verbs. An excerpt of the result can be seen in Figure 3. The only difference to the input file is that the columns for head and dependency relation are filled now. In the third line, we have the word *nun* (now), for example, which is an adverb and stands in modifier relation to its head, the word *komme* (come).

5.6 Extraction of Multiword Expressions

From the parsed corpus, we extracted all verb NP and verb PP combinations where the verb is one of the chosen support verbs: 105,002 tokens and 20,214 types for NPs and 110,413 tokens and 33,560 types for PPs.

For the verb NP combinations, we extracted all verbs and nouns where

[9] http://www.coli.uni-saarland.de/~thorsten/tnt/

1	Ich	_	PPER	PPER	_	2	SB	_	_	
2	komme	_	VVFIN	VVFIN	_	0	ROOT	_	_	
3	nun	_	ADV	ADV	_	2	MO	_	_	
4	zu	_	APPR	APPR	_	2	MO	_	_	
5	den	_	ART	ART	_	4	NK	_	_	
6	vom	_	APPRART		APPRART		_	8	SBP	_	_
7	Ausschuß		_	NN	NN	_	6	NK	_	_
8	für	_	ADJA	ADJA	_	4	NK	_	_	
9	Verkehr		_	NN	NN	_	4	NK	_	_
10	und	_	KON	KON	_	9	CD	_	_	
11	Fremdenverkehr			_	NN	NN	_	9	CJ	_	_
12	eingereichten	_		VVFIN	VVFIN	_	0	ROOT	_	_
13	Änderungsanträgen			_	NN	NN	_	12	SB	_	_
14	:	_	$.	$.	_	12	PUNC	_	_	
15	die	_	ART	ART	_	17	NK	_	_	
16	Sozialdemokratische			_	ADJA	ADJA	_	17	NK	_	_
17	Fraktion		_	NN	NN	_	18	SB	_	_
18	wird	_	VAFIN	VAFIN	_	12	CJ	_	_	
19	mit	_	APPR	APPR	_	26	MO	_	_	
20	Ausnahme		_	NN	NN	_	19	NK	_	_
21	des	_	ART	ART	_	22	NK	_	_	
22	Änderungsantrags			_	NN	NN	_	19	AG	_	_
23	16	_	CARD	CARD	_	25	NK	_	_	
24	sämtliche		_	ADJA	ADJA	_	25	NK	_	_
25	Änderungsanträge			_	NN	NN	_	22	AG	_	_
26	unterstützen	_		VVIZU	VVIZU	_	18	OC	_	_
27	.	_	$.	$.	_	2	PUNC	_	_	

Figure 3: Excerpt of the parser output file in CoNNL format

the noun is labeled as a direct object and its head is the verb. For the verb PP combinations, we took all verbs, prepositions, and nouns where the noun's head is the preposition, and the preposition's head is the verb. We also accounted for relative clauses and *dass* (that) clauses, for which the labeling of the parser seemed to be wrong.

To avoid data sparseness, the potential MWEs were lemmatized with Morphy [10]. Morphy has problems with abbreviations and special signs like "%", so we replaced these tokens. In the end, MWEs occurring less than 50

[10]http://www.wolfganglezius.de/doku.php?id=public:cl:morphy

times were removed, leaving us with 313 verb NP and 231 verb PP combinations.

5.7 Choosing the Best Multiword Expressions

Like Villada Moirón and Tiedemann (2006), we wanted to choose the "best" 200 MWEs to test our method. In contrast to them, we did not combine different measures, but only used the log-likelihood measure to do this. We computed the measure with the NSP Package (Banerjee and Pedersen, 2003). First, we had to run a perl program that counted the frequencies of all the bigrams in the corpus, and then we could measure the log-likelihood ratio for the bigrams. From the resulting rankings, we chose the top 200 verb NP and the top 200 verb PP tuples.

The rankings served as the baseline for our evaluation.

5.8 Creating Rankings

The link lexicon was built by collecting all words with their translations and translation frequencies. An excerpt of the link lexicon can be seen in Figure 4.

```
'folgerung': {'from': 1, 'god': 1, 'suggests': 1, 'likely': 1, 'have': 1, 'follows': 1, 'drawn': 1,
'implication': 1, 'consequently': 1, 'by': 1, 'conclusion': 3},
'prinz': {'king': 1, 'kind': 1, 'takes': 1, 'some': 1, 'able': 1, 'see': 3, 'how': 1, 'finally': 1, 'much': 1,
'senior': 1, 'prince': 15, 'if': 3}
```

Figure 4: Excerpt of the link lexicon

For each verb NP and verb PP tuple, all alignments for the single words in the context of the tuple were collected. An example from this collection for verb PP tuples is shown in Figure 5. The words in the first pair of curly brackets are the words that are aligned to *zu* (to), the words in the second

pair those that are aligned to *Einigung* (agreement), and the words in the third pair those that are aligned to *kommen* (to come). As Villada Moirón and Tiedemann (2006) got the best results without counting NO_LINKS, we left them out as well. This is the reason why the sums of alignment frequencies for the single words are not equal.

'zu Einigung kommen':
[{'be': 4, 'about': 1, 'enable': 1, 'to': 15, 'being': 1, 'come': 6, 'make': 2, 'able': 6, 'use': 1, 'us': 2, 'possible': 2, 'bring': 1, 'made': 1, 'place': 1, 'put': 1, 'impossible': 1, 'opportunity': 1, 'order': 1},
{'conciliation': 2, 'deal': 1, 'conclude': 1, 'agreed': 2, 'find': 2, 'obtain': 2, 'an': 69, 'reached': 21, 'settlement': 1, 'any': 2, 'issues': 1, 'subject': 1, 'no': 1, 'achieved': 1, 'least': 1, 'arrangement': 1, 'compromise': 2, 'agreements': 3, 'between': 1, 'was': 3, 'achieving': 1, 'conclusion': 1, 'sort': 1, 'reaching': 5, 'reach': 64, 'agreement': 122, 'understanding': 2, 'failure': 1, 'on': 12, 'whereby': 1, 'could': 1, 'final': 1, 'consensus': 5, 'facilitate': 1, 'agree': 3},
{'come': 3, 'then': 1, 'led': 1, 'prove': 1, 'had': 1, 'there': 1, 'able': 1, 'it': 1, 'us': 3, 'next': 1, 'will': 2, 'finally': 1, 'let': 1, 'been': 2, 'allow': 1, 'reached': 1, 'arrive': 3, 'was': 2, 'at': 4}]

Figure 5: Excerpt of the collected alignments

After building the link lexicon and the alignment collection, the measures translational entropy and pda could be calculated. We then got four rankings for each language pair: one for NP verb combinations using the entropy measure, one for NP verb combinations using the pda measure, one for PP verb combinations using the entropy measure, and one for PP verb combinations using the pda measure. In contrast to Villada Moirón and Tiedemann (2006) who only analyzed all languages separately, we also combined the values from the different languages pairs by taking their average because we hypothesized that this might lead to better results.

5.9 Evaluation and Results

Six humans annotated the lists of verb NP and verb PP combinations: two experts and four students. They were asked to decide for each expression whether it was an idiom (“i”), an SVC (“s”), or something else (“a”).

The average pairwise **inter-annotator agreement** was 0.673 for NPs and 0.695 for PPs. Since one of the students said she had problems with the annotation and her annotation differed a lot from those of the others, we removed her annotation. As a result, we got a better inter-annotator agreement of 0.7015 for NPs and 0.746 for PPs. The agreement between the two experts was 0.62 for NPs and 0.815 for PPs.

In addition, we calculated the so-called **Fleiss' Kappa** (Fleiss, 1971). We got scores of 0.4689 for verb NP combinations and 0.5758 for verb PP combinations, which are considered "moderate agreement".

We also measured the agreement when counting idioms and SVCs as one class. Then the average pairwise inter-annotator agreement was 0.799 for NPs and 0.845 for PPs, and the Fleiss' Kappa was 0.5866 for NPs ("moderate agreement") and 0.6953 for PPs ("substantial agreement"). The inter-annotator agreement is not so meaningful here because it is clear that the agreement will be higher when having one class less. The Fleiss' Kappa score is meaningful because it considers the number of classes in its formula.

We built a gold-standard out of the five annotations by always taking the majority vote. If there was no majority, we took the decision of one of the experts (who knows more about MWEs than the other expert).

Table 7 shows the uap of the entropy measure and Table 8 the uap of the pda measure for the different MWEs and language pairs. The baseline is the initial ranking of MWEs after applying the log-likelihood ratio. "I" means that we counted only idioms and "I + S" that we counted both idioms and SVCs as correct. The best result for each category is printed in bold.

One can see that the German-English alignment performs well for NP verb combinations, but not so well for PP verb combinations. The German-Finnish alignment seems to be good for PP verb combinations, especially

when using the pda score. In general, the pda score performs much worse than translational entropy. If we count both idioms and SVCs as correct, the results are much better than if we count only idioms as correct. This is not surprising because, as we have shown in chapter 2.4, the distinction between idioms and SVCs is not completely clear, even to humans. The agreement between the annotators also reflects this fact: When counting idioms and SVCs as one class, the scores are much higher than when counting them as separate classes.

For NPs, there is a significant improvement over the baseline when taking the entropy measure and counting both idioms and SVCs as correct. For PPs, the results are under or only slightly over the high baseline.

For the combined ranking, the average of the entropy results (when counting idioms and SVCs as correct) is better than for the single language pairs. This indicates that it is useful to combine the measures from the languages.

Despite our beliefs, the German-Finnish alignment did not bring an improvement compared to the other language pairs. One reason for this might be that the alignment did not work as well as for the other languages. We checked this by calculating the average alignment scores for all language pairs, but found out that the average score for the German-Finnish alignment is the highest. So we cannot be sure if our hypothesis — the method performs better for the alignment between very distant languages — was correct.

In Table 9, one can see an excerpt of the NP ranking for the German-English alignment, using the entropy measure to order the MWEs. Table 10 shows an excerpt of the PP ranking for the German-Dutch alignment, again using the entropy measure. These rankings were chosen as examples because they performed best. One can see that there are almost only idioms and SVCs among the first twenty expressions. Among the expressions in the

		I	I + S
DE-EN	NP	**0.2283**	**0.7678**
	PP	0.3502	0.7415
	Average	0.2893	0.7547
DE-ES	NP	0.1826	0.7094
	PP	**0.4068**	0.8084
	Average	**0.2947**	0.7589
DE-NL	NP	0.2116	0.7129
	PP	0.2660	0.8140
	Average	0.2388	0.7635
DE-FI	NP	0.1386	0.6717
	PP	0.2252	0.6928
	Average	0.1938	0.5940
Combined	NP	0.1969	0.7393
	PP	0.3354	**0.8187**
	Average	0.2662	**0.7790**
Baseline	NP	0.0869	0.5652
	PP	0.4899	0.8083
	Average	0.2884	0.6868

Table 7: Uap of the entropy measure for the different MWE types and languages

		I	I + S
DE-EN	NP	**0.1067**	**0.4558**
	PP	0.1912	0.4769
	Average	0.1490	0.4664
DE-ES	NP	0.0678	0.4229
	PP	0.1555	0.4549
	Average	0.1117	0.4389
DE-NL	NP	0.0714	0.4414
	PP	0.1863	0.5068
	Average	0.1289	0.4741
DE-FI	NP	0.0744	0.4175
	PP	**0.2504**	**0.5728**
	Average	**0.1624**	**0.4952**
Combined	NP	0.0741	0.4414
	PP	0.1810	0.5151
	Average	0.1276	0.4783
Baseline	NP	0.0869	0.5652
	PP	0.4899	0.8083
	Average	0.2884	0.6868

Table 8: Uap of the pda measure for the different MWE types and languages

middle of the ranking, some are idioms, some SVCs, and some "other", and among the expressions in the end, almost all are annotated as "other".

It becomes clear that the method of Villada Moirón and Tiedemann (2006) works well to order expressions according to their idiomaticity, but, like humans, it is not able to distinguish between idioms and SVCs. We have shown that their approach can not only be applied to verb PP, but also to verb NP combinations. For verb PP combinations, the absolute results are better than for verb NP combinations, but the baseline happens to very high. Verb NP combinations outperform their baseline. It is not important which language pair one chooses because the results for all of them are similar, yet it makes sense to combine different language pairs. Only the entropy measure works for German; the pda score performs very badly.

Rank	MWE	Entropy	Pda	Annotation
1	Wert legen	3.3736	0.0022	s
2	Verwirrung stiften	3.2581	0.0059	s
3	Weg gehen	3.2050	0.0035	i
4	Zustimmung geben	3.1141	0.0033	s
5	Antrag stellen	3.1078	0.0041	s
6	Gedanken machen	3.0668	0.0052	s
7	Vorzug geben	3.0292	0.0132	s
8	Bilanz ziehen	2.9964	0.0047	s
9	Auskunft geben	2.9592	0.0099	s
10	Nachdruck legen	2.9389	0.0027	s
11	Schwerpunkt legen	2.8238	0.0071	i
12	Forderung stellen	2.8149	0.0057	s
13	Grundstein legen	2.7954	0.0061	i
14	Vorwurf machen	2.7660	0.0213	s
15	Mittel stellen	2.7293	0.0044	a
16	Frage stellen	2.6861	0.0003	s
17	Recht stellen	2.6808	0.0075	a
18	Weichen stellen	2.6807	0.0000	i
19	Frage kommen	2.6473	0.0039	a
20	Antwort geben	2.6458	0.0015	s
101	Widerstand geben	1.8631	0.0143	a
102	Zusammenhang geben	1.8500	0.0047	a
103	Rat machen	1.8460	0.0032	a
104	Finger legen	1.8452	0.0099	a
105	Anfrage stellen	1.8431	0.0103	s
106	Sicherheit geben	1.8429	0.0046	s
107	Verfahren geben	1.8162	0.0077	a
108	Art geben	1.8131	0.0065	a
109	Bereich geben	1.8027	0.0022	a
110	Ausnahme machen	1.7934	0.0127	s
111	Spielraum geben	1.7894	0.0096	i
112	Geld geben	1.7648	0.0067	a
113	Abstimmung geben	1.7622	0.0074	a
114	Debatte geben	1.7491	0.0069	a
115	Instrument geben	1.7400	0.0091	a
116	Auffassung geben	1.7287	0.0185	a
117	Ansicht geben	1.7286	0.0100	a
118	Verbesserung geben	1.7255	0.0156	a
119	Gebiet geben	1.7127	0.0063	a
120	Mehrheit geben	1.7126	0.0067	a
181	Regierung machen	1.2771	0.0096	a
182	Leute geben	1.2347	0.0072	a
183	Dialog geben	1.2310	0.0112	a
184	System geben	1.2035	0.0105	a
185	Land machen	1.1953	0.0122	a
186	Bericht geben	1.1889	0.0025	s
187	Richtlinie geben	1.1826	0.0063	a
188	Unterschied geben	1.1746	0.0099	a
189	Abgeordnete machen	1.1525	0.0213	a
190	Freiheit geben	1.1479	0.0149	s
191	Land kommen	1.0980	0.0045	a
192	Element geben	1.0928	0.0270	a
193	Frieden geben	1.0888	0.0061	a
194	Problem geben	1.0795	0.0022	a
195	Demokratie geben	1.0747	0.0068	a
196	Gemeinschaft geben	1.0551	0.0075	a
197	Markt geben	0.9737	0.0116	a
198	Kollege geben	0.9473	0.1111	a
199	Gruppe geben	0.8117	0.0154	a
200	Programm geben	0.7624	0.0102	a

Table 9: Ranking for NP verb tuples and the DE-EN alignment, using the entropy measure

Rank	MWE	Entropy	Pda	Annotation
1	zu eigen machen	3.1173	0.0079	s
2	zum Tragen kommen	3.0731	0.0172	i
3	zum Ausdruck kommen	3.0263	0.0025	s
4	ans Herz legen	2.9379	0.0150	i
5	zum Schluss kommen	2.8240	0.0054	s
6	zu Protokoll geben	2.7767	0.0086	s
7	zu Ende gehen	2.7457	0.0090	s
8	zur Kenntnis geben	2.6083	0.0270	s
9	zu Ergebnis kommen	2.6000	0.0052	s
10	in Erinnerung rufen	2.5481	0.0039	s
11	ins Detail gehen	2.5465	0.0090	s
12	zum Opfer fallen	2.4911	0.0103	s
13	zum Einsatz kommen	2.4783	0.0207	s
14	zur Anwendung kommen	2.4463	0.0062	s
15	vor Gericht stellen	2.4324	0.0069	s
16	zu Einigung kommen	2.4287	0.0072	s
17	unter Beweis stellen	2.4109	0.0032	s
18	zu Schluss kommen	2.3802	0.0077	s
19	zu Wort kommen	2.3799	0.0090	i
20	zur Sprache kommen	2.3707	0.0043	s
101	in Aktion treten	1.4605	0.0508	s
102	um Dinge gehen	1.4598	0.0195	a
103	um Umsetzung gehen	1.4570	0.0138	a
104	in Auftrag geben	1.4565	0.0132	s
105	auf Tisch kommen	1.4556	0.0204	i
106	an Arbeit machen	1.4407	0.0191	s
107	um Punkt gehen	1.4257	0.0370	a
108	in Parlament kommen	1.4161	0.0133	a
109	in Weg legen	1.4157	0.0349	i
110	bei Frage gehen	1.4069	0.0333	a
111	in Bereich machen	1.4030	0.0160	a
112	um Entscheidung gehen	1.4026	0.0282	a
113	aus Grund geben	1.4009	0.0652	a
114	in Bereich gehen	1.4008	0.0159	a
115	in Rechnung stellen	1.3958	0.0316	s
116	ins Parlament kommen	1.3864	0.0169	a
117	in Union kommen	1.3828	0.0077	a
118	in Vergangenheit geben	1.3790	0.0109	a
119	aus Verkehr ziehen	1.3775	0.0123	i
120	im Bericht kommen	1.3727	0.0149	a
181	aus Land kommen	0.9443	0.0044	a
182	in Mitgliedstaaten geben	0.9402	0.0048	a
183	in Jahr geben	0.9242	0.0455	a
184	in Welt geben	0.9065	0.0104	a
185	um Geld gehen	0.8918	0.0173	a
186	um Interesse gehen	0.8741	0.0909	a
187	um Menschenrechte gehen	0.8710	0.0167	a
188	in Jahr kommen	0.8322	0.0625	a
189	in Land geben	0.8223	0.0076	a
190	um Entwicklung gehen	0.8043	0.0238	a
191	auf Markt kommen	0.7991	0.0120	s
192	in Wettbewerb treten	0.7824	0.0400	s
193	um Zusammenarbeit gehen	0.7737	0.0250	a
194	um Sicherheit gehen	0.7685	0.0076	a
195	um Mensch gehen	0.6562	0.2222	a
196	um Recht gehen	0.5869	0.0138	a
197	in Jahr machen	0.5855	0.0857	a
198	unter Richtlinie fallen	0.4379	0.0244	a
199	um Land gehen	0.3890	0.0667	a
200	seit Jahr geben	0.0000	0.2500	a

Table 10: Ranking for PP verb tuples and the DE-NL alignment, using the entropy measure

6 Token-based Approach

The approach by Villada Moirón and Tiedemann (2006) is type-based: It ranks MWEs according to their idiomaticity, but does not take into account that an MWE could be used both literally and idiomatically in different contexts.

One of our goals was to extend their approach in such a way to make it token-based. Our method should be able to decide for the single occurrences of an MWE in context whether it is used idiomatically or literally.

6.1 Idea

The idiomatic and literal usages of an MWE are like two different word senses. As clustering has been applied to word sense discrimination before (Purandare and Pedersen, 2004), this should also be useful for our task.

The input to our algorithm is an occurrence of an MWE in context, and the question is if the MWE is used idiomatically or literally. All occurrences of the MWE together with their contexts are extracted from the corpus. These occurrences, plus the occurrence in question, are then divided into two clusters by representing them as context vectors and calculating the similarity between them. For each of the clusters, we can measure its average translational entropy by collecting alignments and calculating the measure like in chapter 4.3. The cluster with the higher translational entropy is the idiomatic cluster and the other one the literal cluster. If the occurrence in question is found in the idiomatic cluster, it is classified as idiomatic; otherwise, it is classified as literal.

6.2 Extracting Contexts

For the given MWE in context, we extracted all of its occurrences from the Europarl corpus, together with 30 words to the left and 30 words to the right. We added the given MWE with its context to the list. Then we used the program txt2sval.pl (contained in SenseClusters [11]) to convert the file into the *sval format* which the SenseClusters clustering algorithm needs as input. An excerpt of an sval file can be seen in 6. The target word (*Rolle spielen* — to play a role) is marked with "<head>" tags.

```
<instance id="3">
<answer instance="3" senseid="NOTAG"/>
<context>
außerhalb des agrarsektors zu unternehmen unter anderem in den bereichen ländlicher tourismus
sport kultur sanierung der ressourcen umstellung von unternehmen neue technologien
dienstleistungen usw. doch obwohl die landwirtschaft keine ausschließliche
<head>rolle_spielen</head> mehr ist sie weiterhin wichtig nicht nur um den wirtschaftlichen und
sozialen niedergang der ländlichen gebiete und das entstehen von weiteren geisterstädten zu
verhindern sondern auch weil den landwirten eine
</context>
</instance>
<instance id="4">
<answer instance="4" senseid="NOTAG"/>
<context>
ein paar kurze worte zu sagen um zwei punkte hervorzuheben an die uns diese berichte erinnern
und die in der vorstellung die wir von der union haben eine wesentliche strategische
<head>rolle_spielen</head> der erste ist die maßgebliche und zentrale bedeutung die wir dem
grundsatz des wirtschaftlichen und sozialen zusammenhalts nach wie vor beimessen wir geben
unserer besorgnis über meldungen ausdruck denen wir
</context>
</instance>
```

Figure 6: Excerpt of a context file in sval format

6.3 Clustering

SenseClusters tries to cluster input contexts in order to discover the different meanings of a target word. It is based on the idea that words that occur in

[11] http://www.d.umn.edu/~tpederse/senseclusters.html

similar contexts will have similar meanings. The various similar contexts in which a target word occurs will reflect its different meanings.

Context discrimination is performed by using either first or second order representation. **First order representation** means that for each context a vector is created that indicates which features (unigrams, bigrams, co-occurrences, or target co-occurrences) occur in that context. **Second order representation** means that for each context a vector is created that indicates which words occur with the words in that context (second order co-occurrences). Purandare and Pedersen (2004) claim that first order representation works better for large amounts of data and second order representation for small amounts of data.

SenseClusters implements various clustering methods, out of which we chose Repeated Bisections. This is a hybrid approach, combining hierarchical and partitional methods.

Hierarchical approaches build (*agglomerative*) or break up (*divisive*) a hierarchy of clusters.

Partitional approaches, in contrast, divide an entire set of instances into a predetermined number of clusters without going through a series of pairwise comparisons. A well-known partitional approach is **K-means clustering**. In this approach, K random vectors are selected to serve as centroids of these initial K clusters. Every other vector is assigned to one of the K clusters whose centroid is closest to that vector. After all vectors are assigned, the cluster centroids are recomputed by averaging all of the vectors assigned to that cluster. This is repeated until no vector changes its cluster anymore.

Repeated Bisections overcomes the main weakness with partitional approaches: their instability in clustering due to the choice of the initial random

centroids. It starts with all instances in a single cluster. At each iteration, one cluster is selected whose bisection optimizes the chosen criteria function. A cluster is bisected using the standard K-means method with $K = 2$.

We used the web interface of SenseClusters [12] to cluster our instances. The options we chose were:

- SenseClusters Native Methodology
- Target Word Clustering
- Cluster Stopping: Set manually
- Statistical Test of Association: Log-likelihood
- Number of Clusters: 2

We kept the default settings for all other options. For details of these options, see the web interface of SenseClusters.

6.4 Collecting Alignments and Calculating Translational Entropy

We collected all alignments for the words of the MWE in the context of the MWE, using the German-English alignment. Then we calculated the average translational entropy for each of the two clusters as described in chapter 4.3. The cluster with the higher average translational entropy was the idiomatic and the other one the literal cluster.

[12] http://marimba.d.umn.edu/cgi-bin/SC-cgi/index.cgi

6.5 Evaluation and Results

Due to technical reasons, we could unfortunately only evaluate a very small number of MWEs: five MWEs with four tokens each. We took the tokens in context from the ECI corpus [13], the Gutenberg corpus [14], and the internet. The tokens were annotated by a human expert as idiomatic or literal. Our method was applied and gave the results in Table 11.

In the columns "literal" and "idiomatic", one can see which classifications have been predicted by our program and how many of the predicted classifications are correct. The "accuracy" column gives the accuracy for each MWE, and the "avg. acc." column on the right gives the average accuracy for MWEs of the form PP verb and MWEs of the form NP verb. We also show the average accuracies for literal and idiomatic classification. At the bottom, one can see the average accuracy for all MWEs. The baseline is always predicting the idiomatic label.

The results for PP verb combinations are better than for NP verb combinations, but both types of MWEs do not outperform the baseline. We cannot conclude that the presented method works. Yet if one looks at the data, it becomes clear why the results have to be bad: In the Europarl corpus, potential idioms are mainly (some only) used idiomatically, that is why it is not possible to get two correct clusters, one representing the literal and one the idiomatic usage of an idiom. As an example, we analyzed all 205 extracted contexts of the idiom *Schritt halten* and found that none of these occurrences is literal.

We searched for a different parallel corpus, but did not find any that was big enough for the web interface to work. The web interface needs a certain

[13] http://www.elsnet.org/eci.html

[14] http://gutenberg.spiegel.de/

	MWE	Literal		Idiomatic		Accuracy	Avg. Acc.
		pred.	correct	pred.	correct		
PP	auf Beine stellen	1	0	3	1	0.25	0.50
	in Hand haben	3	2	1	1	0.75	
NP	Rolle spielen	4	2	0	0	0.50	0.42
	Schritt halten	4	2	0	0	0.50	
	Weg ebnen	1	0	3	1	0.25	
	Avg. Accuracy	0.60		0.30			
	Accuracy	0.45					
	Baseline	0.50					

Table 11: Accuracy for the classification of different MWE tokens

size of the input file to be able to find enough features.

Another reason why our approach did not work could be that we took the tokens to be classified from different sources than Europarl, that means, we mixed different domains. We did that because it was not feasible to search for the very few literal occurrences of the tokens in the Europarl corpus.

7 Conclusion

The approach of Villada Moirón and Tiedemann (2006) is based on the hypothesis that one can discover idioms by looking at their alignments to other languages. The more diverse an alignment is and the less words are translated like their default translations, the more idiomatic an expression is. We have shown that this approach works and can also be applied to German MWEs. We have implemented several extensions, some of which were successful and some were not:

Extension	Success
NP verb combinations	Yes
Alignment to Finnish	No
Combination of languages	Yes
Token-based approach	No

The approach of Villada Moirón and Tiedemann (2006) works well to order expressions according to their idiomaticity, but it is not able to distinguish between idioms and SVCs. This reflects the fact that this distinction is difficult and unclear in general — even humans have problems with it.

Although the German-Finnish alignment did not bring an improvement, the hypothesis that distant languages perform better might be worth further research. One could try to work with other non-Indo-European languages, for example Hungarian (which is not included in the Europarl corpus).

For future work, it would also be interesting to further explore the token-based approach. It might be the case that when using a different parallel corpus or when not mixing different domains, the results would be much better.

References

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., Hansen, S., Hirschmann, H., Janitzek, J., Kirstein, C., Langner, R., Michelbacher, L., Plaehn, O., Preis, C., Pußel, M., Rower, M., Schrader, B., Schwartz, A., Smith, G., and Uszkoreit, H. (2003). Tiger Annotationsschema.

Bahr, B. I. (1977). *Untersuchungen zu Typen von Funktionsverbfügungen und ihrer Abgrenzung gegen andere Arten der Nominalverbindung*. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, number 18, pages 89–96, Morristown, NJ, USA. Association for Computational Linguistics.

Banerjee, S. and Pedersen, T. (2003). The Design, Implementation, and Use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City.

Bannard, C. (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, Czech Republic. Association for Computational Linguistics.

Bannard, C., Baldwin, T., and Lascarides, A. (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, Morristown, NJ, USA. Association for Computational Linguistics.

Birke, J. and Sarkar, A. (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.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of LREC 2006*.

Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Daniels, K. (1963). *Substantivierungstendenzen in der deutschen Gegenwartssprache. Nominaler Ausbau des verbalen Denkkreises*. Düsseldorf.

Fazly, A. and Stevenson, S. (2007). Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382.

Higi-Wydler, M. (1989). *Zur Übersetzung von Idiomen. Eine Beschreibung und Klassifizierung deutscher Idiome und ihrer französischen Übersetzungen*, volume 146 of *Europäische Hochschulschriften, Reihe XIII, Französische Sprache und Literatur*. Peter Lang.

Katz, G. and Giesbrecht, E. (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, Australia. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit 2005*.

Kuhn, J. (1994). The treatment of support verb constructions in HPSG-based Machine Translation.

Lin, D. (1999). Automatic identification of noncompositional phrases. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 317–324.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

McCarthy, D., Keller, B., and Carroll, J. (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.

Melamed, I. D. (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)*.

Melamed, I. D. (1997b). Measuring Semantic Entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 41–46.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*.

Purandare, A. and Pedersen, T. (2004). Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In H. T. Ng and E. Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 41–48, Boston, Massachusetts, USA. Association for Computational Linguistics.

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

Sag, I. A., Baldwin, T., Bond, F., and Copestake, A. (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-2002)*, pages 1–15.

Venkatapathy, S. and Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906, Morristown, NJ, USA. Association for Computational Linguistics.

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *EACL Workshop On Multi-Word-Expressions In A Multilingual Context*.

Wermter, J. and Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 843–850, Morristown, NJ, USA. Association for Computational Linguistics.