Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b, 70569 Stuttgart

A Cross-Lingual Approach to Metaphor Identification

Max Kisselew
Matrikelnummer: 2352413
E-Mail: max.kisselew@ims.uni-stuttgart.de

Diplomarbeit Nr. 121

Begonnen am: 01. April 2012
Beendet am: 13. September 2012

vorgelegt bei
Prüfer: Prof. Dr. Hinrich Schütze
Betreuer: Christian Scheible

Contents

1 Introduction

Figurative expressions like "dark humor", "deep respect" or "to surf the web" are so common in everyday use that speakers mostly do not even notice how they use them or how these and similar expressions are used by others. Metaphors are usually used to describe emotions, attitudes or abstract concepts in terms of common physical experiences or similar, more concrete concepts to make the former more comprehensible. Thus, metaphors arise when a similarity between two concepts is established. For example, pleasant feelings often cause a sensation of warmth and vice-versa. Therefore, we often describe a handshake, an applause or a relationship as *warm* because we associate pleasant feelings with it.

According to Lakoff and Johnson (1980) metaphors are often constructed by applying an expression from a concrete domain to a more abstract concept. For example, human relationships can be described using temperature or sensation expressions:

Literal:	*warm beer*
Metaphorical:	*warm relationship*

By translating this example to German, we get the following expressions:

Literal:	*warmes Bier*
Metaphorical:	**warme Beziehung*

Warme Beziehung is quite uncommon in German. Instead, expressions like *herzliche Beziehung* (*cordial relationship*) or *warmherzige Beziehung* (*warmhearted relationship*) are more appropriate. This example demonstrates that metaphors cannot always be translated from one language to another language without any adaptation.

This observation leads to our hypothesis:

While the direct translation of a literal expression to another language will be mostly acceptable, the direct translation of a metaphorical expression will mostly fail.

We say "mostly" because we do not expect all literal and metaphorical expressions to follow this assumption. Literal expressions are often conventionalized

or have different naming reasons. For instance, *harte Süßigkeit*, the direct German translation for *hard candy* does not exist in German. The correct translation would be *Bonbon*. But metaphorical expressions, can be used similarly across languages as well. Especially the subgroup of *conventional metaphors* which comprises frequently used expressions with a metaphorical origin. An example for a conventional metaphor is "to grasp a theory" because the verb *to grasp* was primarily used for physical objects and not for abstract entities like theories etc. Other conventional metaphors describe emotions in terms of properties of things and are used across languages in the same way. For example the adjective *deep* which is frequently used to describe the intensity of emotions:

	English	German
Literal:	*deep gorge*	*tiefe Schlucht*
Metaphorical:	*deep sadness*	*tiefe Trauer*

We expect that due to their frequent everyday use conventional metaphors behave much like literal expressions and therefore occur in other languages as well.

Metaphor identification, or metaphor recognition, is usually distinguished from metaphor interpretation. While the former describes the classification of an expression into being either literal or non-literal, the latter refers to the identification of "the intended literal meaning of a metaphorical expression" (Shutova et al., 2012). This thesis treats only the metaphor identification task because it is still a task in natural language processing far form being accomplished perfectly. Chapter 5 presents an overview of the related work in this field.

The aim of this thesis is to verify the hypothesis formulated above. We first devise a method for feature extraction from literal and metaphorical adjective-noun phrases and clustering them. Out method is not restricted to any domain and does not rely on specific a priori knowledge. For this purpose we translate a set of literal and metaphorical English adjective-noun phrases into German, French, Spanish, Estonian and Russian using on-line plain text dictionaries. Then we calculate the frequencies and other statistical measures by searching the corpora for the language of the respective translation. The features extracted this way are converted to feature vectors for each adjective-noun phrase. Thus we use the vector space model (Salton et al., 1975) to separate feature vectors originating from literal and metaphorical adjective-noun phrases. The separation itself is carried out automatically by the K-means clustering algorithm. The method is described in more detail in Chapter 2. Then we implement a system for the verification of this method. The data used for our implementation and our experiments are shown in Chapter 3. The conducted experiments and their results are given in Chapter 4. A conclusion and perspectives for future work are presented in Chapter 6.

2 Method

Our hypothesis from Chapter 1 states that translations of metaphorical expressions are expected not to occur at all or to occur less often in the appropriate corpora than translations of literal expressions. To verify this hypothesis we translate adjective-noun phrases from a source language to other languages. We then search for the occurrences of the translated phrases in corpora compiled from data for these languages and compute statistical features for them. These features are then used to build feature vectors which are clustered using the K-means clustering algorithm. The expectation is that the instances will be divided into two clusters: A literal cluster and a metaphorical cluster. The approach is illustrated in Figure 2.1. In the following sections we explain the steps of our method in more detail. The data we use for our implementation of the method are described in Chapter 3.

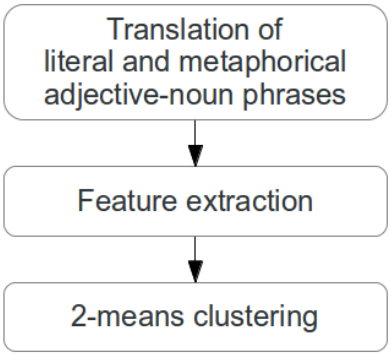

Figure 2.1: Approach of our method

2.1 Translation of Adjective-Noun Phrases

As a starting point we use a set of literal and metaphorical adjective-noun phrases such as *deep bowl* (literal) and *deep affection* (metaphorical) in English. These phrases are translated automatically to different target languages. For example, the entries for the English words *deep*, *bowl* and *affection* from the English-German dictionary are shown in Table 2.1.

English	German
deep	tief, unergründlich, dunkel, tiefgehend, Kolk
snow	Bildrauschen, verrauschtes Bild, Schnee, Ameisenkrieg, Koks
affection	Affektion, Gunst, Wohlwollen, Zuneigung

Table 2.1: English-German dictionary entries for *deep*, *bowl* and *affection*

When translating a sequence of words from a source language, all available translations into the target language are combined with each other. Consider an adjective-noun phrase in a particular source language. Since a phrase consists of two words we can refer to the first word as s_1 and to the second word as s_2. There are m possible translations into a target language for the first word s_1 and n potential translations of the second word s_2. To retrieve all possible translations of the source language phrase we therefore combine all m translations for word s_1 with the translations for word s_2 by means of the algorithm illustrated in Figure 2.2. After the execution of the algorithm we get a list of all possible translations of a source language adjective-noun phrase into phrases of a target language.

```
translations_for_w1 <- get_translations_from_dictionary(w1)
translations_for_w2 <- get_translations_from_dictionary(w2)
translation_candidates <- []

do for i <- 1 to m:
  do for j in 1 to n:
     translation <- translations_for_w1[i] + translations_for_w1[j]
     translation_candidates.append(translation)
return translation_candidates
```

Figure 2.2: Algorithm for the retrieval of translation candidates into a particluar target language.

The list of translations constructed by the algorithm shown in Figure 2.2 contains a large amount of inaccurate translations. Thus a first filter removes translations which consist of more than one word. So in the example from Table 2.1 the translation *verrauschtes Bild* is removed before combining the remaining translations. By this means the German translations shown in Table 2.2 are produced for the English phrase *deep snow*.

The translations shown in Table 2.2 still contain a large amount of implausible phrases. By implausible phrases we mean sequences of words which cannot occur as translations for adjective-noun phrases. For example, when translating from English to German, it is very unlikely that an English adjective-noun phrase will be translated into a sequence of two adjectives or two nouns. So the implausible

Noun / Adjective	**Bildrauschen**	**Schnee**	**Ameisenkrieg**	**Koks**
tief	tief Bildrauschen	tief Schnee	tief Ameisenkrieg	tief Koks
unergründlich	unergründlich Bildrauschen	unergründlich Schnee	unergründlich Ameisenkrieg	unergründlich Koks
dunkel	dunkel Bildrauschen	dunkel Schnee	dunkel Ameisenkrieg	dunkel Koks
tiefgehend	tiefgehend Bildrauschen	tiefgehend Schnee	tiefgehend Ameisenkrieg	tiefgehend Koks
Kolk	Kolk Bildrauschen	Kolk Schnee	Kolk Ameisenkrieg	Kolk Koks

Table 2.2: German translations for the English phrase *deep snow* after removing translations consisting of more than one word.

Noun / Adjective	**Bildrauschen**	**Schnee**	**Koks**
tief	tief Bildrauschen	tief Schnee	tief Koks
unergründlich	unergründlich Bildrauschen	unergründlich Schnee	unergründlich Koks
dunkel	dunkel Bildrauschen	dunkel Schnee	dunkel Koks
tiefgehend	tiefgehend Bildrauschen	tiefgehend Schnee	tiefgehend Koks

Table 2.3: German translations for the English phrase *deep snow* after removing implausible translations.

phrases should not be confused with wrong translations.

In order to remove the implausible translations another two-step filter is applied. In the first step all translations are translated back to the original source language. Those translations which cannot be translated back are removed. In the second step those translations are removed where at least one word does not occur in our corpora. For the example above the following translations are removed: *Kolk Koks*, *Kolk Schnee*, *tief Ameisenkrieg*, *Kolk Ameisenkrieg*, *Kolk Bildrauschen*, *dunkel Ameisenkrieg*, *tiefgehend Ameisenkrieg*, *unergründlich Ameisenkrieg*. After applying the filters we get the translations shown in Table 2.3. Thus the filtering steps help to remove irrelevant and implausible translations.

In the subsequent experiments only one translation for a particular source language adjective-noun phrase is chosen. This is described in more detail in the next section.

2.2 Feature Extraction

We apply a corpus-based feature extraction approach. This means above all that we search for source language adjective-noun phrases and their translation candidates into several target languages in appropriate text corpora. The corpora should comprise a similar content and not differ too much in size. Otherwise the outcoming results could be potentially biased. Two types of corpora fulfill these requirements: Comparable corpora and parallel corpora. We use both kinds of corpora since we also want to find out to which extent the choice of the corpora affects the clustering of literal and metaphorical adjective-noun phrases.

The corpora are tokenized, lemmatized and annotated with part-of-speech tags. Given these annotations we extract a list of adjective-noun phrases for each language from the corpora. An example for German is given in Table 2.4. We then search for translations of the source language adjective-noun phrases in the obtained adjective-noun lists. For each translation the features shown in Table 2.5 are extracted with respect to the corpora of the respective language. They can be subdivided into three groups:

1. **Frequencies of the adjective (*freq-adj*), the noun (*freq-nn*) and the entire phrase (*freq*).**
   These features capture the raw frequency of the adjective, the noun and the entire adjective-noun phrase. *freq-adj* counts, how often the adjective in the translation occurs in the list of adjective-noun phrases for the respective language. In the same manner *freq-nn* is the count of occurrences of the noun in the translation. These two features are included to examine whether the frequency of an adjective or noun and its metaphorical usage correlate. *freq*, in turn, is the frequency of the entire translated adjective-noun phrase. This

Adjective	Noun
@ord@	Dezember
unterbrochen	Sitzungsperiode
europäisch	Parlament
schön	Ferien
schrecklich	Naturkatastrophe
nah	Tag
verschieden	Land
europäisch	Union
...	...

Table 2.4: Adjective-noun pairs extracted from German parallel corpora

Feature	Description
freq-adj	Frequency of the adjective w.r.t. all adjective-noun phrases.
freq-nn	Frequency of the noun w.r.t. all adjective-noun phrases.
freq	Frequency of the adjective-noun tuple w.r.t. all adjective-noun phrases.
ADJ-nn	Number of different adjectives which occur together with the noun.
adj-NN	Number of different nouns which occur together with the adjective.
PMI	Pointwise Mutual Information of the phrase.
χ^2	Pearson's chi-square test for the adjective-noun phrase.

Table 2.5: Features of adjective-noun phrase translations.

is a central feature for the verification of our initially formulated hypothesis. It captures whether the translated adjective-noun phrase has a similar or a different frequency compared to the original source language phrase.

2. **How general, respectively, how versatile is the adjective/noun? (*ADJ-nn*, *adj-NN*)**
   ADJ-nn captures the amount of different adjectives, the noun can occur with. In the same way *adj-NN* is the number of different nouns the adjective can occur with. These features are included to examine whether words appearing in many contexts tend to be used metaphorically more often or not.

3. **Is the adjective-noun-phrase a potential collocation or is it composed by chance? (*PMI*, χ^2)**
   Association measures like the Pointwise Mutual Information (*PMI*) or the χ^2 *test* (chi-square test) are used to calculate whether two words occur together just by chance or whether there is a special kind of association between them which makes them co-occur more frequently. By means of these two features we want to examine whether literal and metaphorical phrases can be reliably separated by associational strength. We expect that metaphorical adjective-

noun phrases have lower association scores than literal adjective-noun phrases due to the diversity of the former.
The Pointwise Mutual Information (PMI) is an information-theoretical "measure of association between elements" (Manning and Schütze, 1999). It tells how likely two words (x and y) will occur together and is mathematically defined as follows:

$$PMI(x, y) = log \frac{p(x, y)}{p(x)p(y)}$$

We assume that $p(x)$ and $p(y)$ correspond to the raw frequencies of the adjective (*freq-adj*) and the noun (*freq-nn*).
Pearson's chi-square (χ^2) is used to calculate to which extent the observed frequency of the adjective-noun phrase differs from the frequency expected for the independence of the adjective and the noun. For this purpose the chi-square test makes use of the null hypothesis which assumes the independence of the adjective and the noun. The greater the difference between the expected and the observed frequency of the adjective-noun phrase, the higher the probability that the null hypothesis can be rejected. See Manning and Schütze (1999) for a more detailed explanation of the χ^2 test.

The features are first extracted for the source language phrases as well as for each translation into a target language resulting in a list of feature vectors. Then the feature vectors of each language are concatenated. This process is visualized in Figure 2.3. The variables e, g, f and s stand for the feature values. Each variable refers to the feature values of a particular target language. For example, e stands for the feature values of the English feature vector which is assumed to be the source language, g refers to feature values from the feature vector for German etc. The first index of a variable denotes the number of an adjective-noun phrase while the second index indicates the feature number.

Thus the concatenated feature vectors consist of feature values for the original source language adjective-noun phrase as well as for the feature values of its translations. However, we saw in the preceding section that, for example, translating an English phrase into other languages results in several translation alternatives. Therefore, one translation from these alternatives has to be selected. This is due to two reasons. First, we want to avoid that particular adjective-noun phrases get a higher influence during the clustering process by virtue of their higher number of translations. Second, the translation process yields a different amount of translations for each language which would make a concatenation of feature vectors impossible. Therefore, we apply four different approaches explained below to get one translation feature vector for an adjective-noun phrase. The subsequent experiments are conducted separately using feature vectors generated by all four

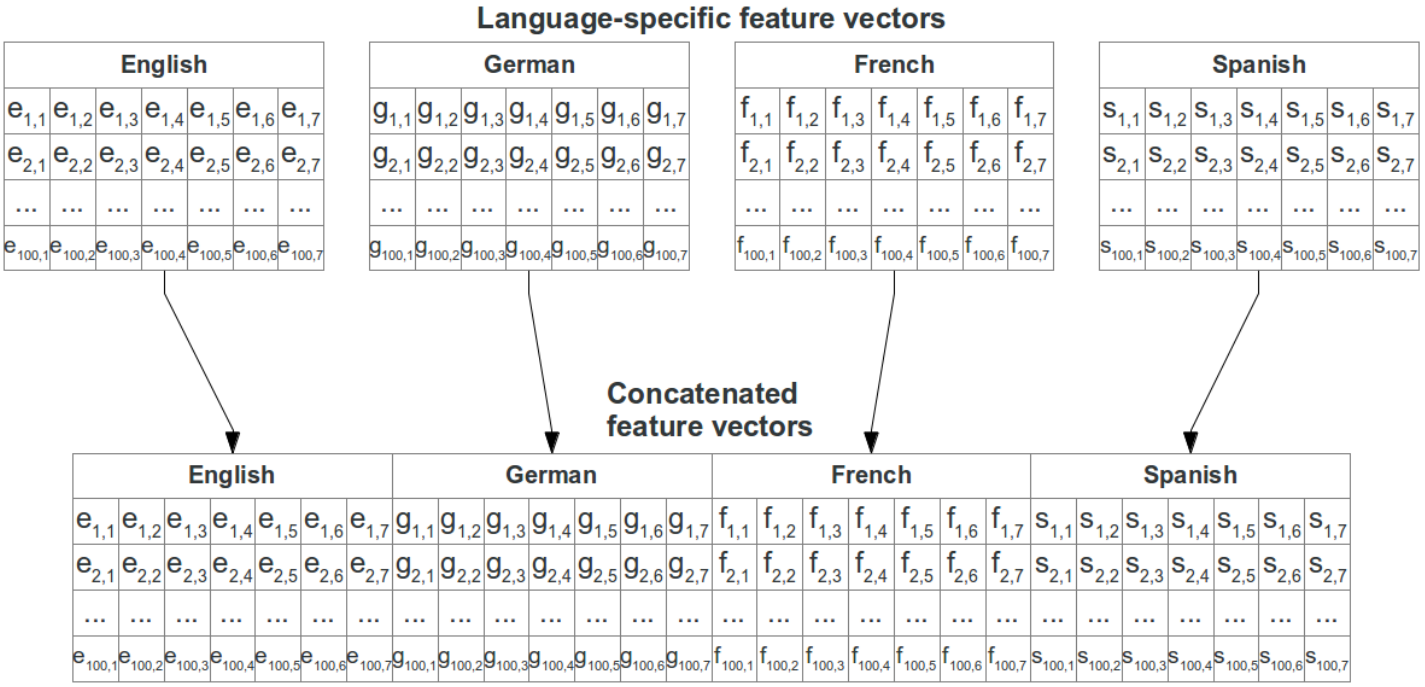

Figure 2.3: Process of feature vectors concatenation

approaches.

The four approaches for getting one translation candidate can be subdivided into two types. The first type applies a simple heuristic to select the best translation candidate from all translations while the second type calculates averages of the translations' feature values. Both methods are described in more detail below.

The first approach we use to select one translation candidate from all translations of a source language adjective-noun phrase employs the following heuristic: Only the translated phrase with the highest frequency (with the highest *freq* feature value) is selected as the best translation candidate. The feature vector of this translation is then included in the concatenated feature vector for a particular source language adjective-noun phrase as shown in Figure 2.3 where English is considered the source language. However, it may happen that among the translations there is no translation which actually occurs in a corpus. Then all *freq* feature values of the translations would be 0.0. In this case the translation with the highest adjective and noun frequencies (with the highest *freq-adj* and *freq-nn* feature values) is selected as the best translation candidate. We refer to the experiments using the feature vector constructed by this approach as "Best-Translation".

The first approach for the selection of a best translation candidate presented above

cannot ensure that the best translation is always selected. In particular, it might happen that a metaphorical adjective-noun phrase is translated to a phrase which preserves the original metaphorical meaning in the target language. Therefore, we use a second method to find a candidate translation for an source language adjective-noun phrase. It works as follows: For a particular feature, all feature values of the translations of the original source language phrase are averaged. These averages are calculated for every feature and constitute the feature values of the translation candidate feature vector. The motivation for the averaging is the expectation that the amount of different translations implicitly captures the nature of the adjective-noun phrase and might therefore help to separate literal from metaphorical phrases.

The averages are calculated in three alternative ways resulting in three types of feature vectors. First, the values are averaged. We name the experiments using this feature vector type "Mean". Second, the median of the values is calculated. By doing this we try to minimize the influence of extreme feature values. We refer to these experiments as "Median". The third way to calculate the averages is to compute the standard deviation of the translations' feature values. The assumption is that the standard deviation of the translations' values might implicitly capture the nature of an adjective-noun phrase similar to a fingerprint what could help to improve the separation of literal and metaphorical vectors. Experiments based on this type of feature vector are indicated by "Std".

An additional optional modification of the averaged data excludes zero values from the calculation of the three average measures. We add the label "w/o 0" to indicate the experiments where the underlying data has been computed this way.

Feature Vectors

As explained in the previous section, a concatenated feature vector is constructed from single feature vectors originating from the translations of a source language adjective-noun phrase into different languages. Due to potential differences in size of the corpora and the requirements of the K-means clustering algorithm the feature values have to be normalized by two steps which are presented below.

In the first normalization step feature values of every feature (values of each column of the concatenated feature vector) are mapped to a scale between 0 and 1. This is carried out to avoid that particular features unintentionally get a higher weight due to higher values. This linear normalization is performed by means of the following formula:

$$x_{normalized} = \frac{x - min}{max - min}$$

In the formula x is the original feature value, min the minimum feature value of the feature and max the maximum value of the feature. The K-means algorithm

requires the feature vectors to have unit length. In a second normalization step we therefore normalize each feature vector to a length of 1.0 by dividing each of its values by the vector's length. The length of an n-dimensional vector $\vec{x}$ is calculated by means of the following formula:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Note that the first normalization step performs a column normalization while the second normalization step normalizes the row values. Now the feature vectors can be used as input vectors for the K-means algorithm.

2.3 Clustering

We want to find out whether literal or metaphorical adjective-noun phrases have similar feature values across languages. Therefore, having the concatenated feature vectors as a basis we can make use of the vector space model (Salton et al., 1975) and apply a clustering algorithm in order to partition the feature vectors into clusters according to their properties. K is the parameter which denotes how many clusters are to be found. Since we intend to separate feature vectors into a literal and a metaphorical cluster, 2-means clustering is carried out.

Classifying approaches are usually subdivided into two categories: supervised and unsupervised approaches. Supervised methods "learn" from annotated training data and classify new unlabeled data by comparing it to the learned training data. Unsupervised approaches, in turn, operate on unlabeled data and separate the data into classes based on the features of the data. Clustering approaches are usually unsupervised which means they do not rely on any training data to perform the classification. The data is assigned to the clusters simply by means of the features of the data. The reason why we opt for a clustering algorithm is that our test set is not sufficiently large to provide a reliable data base for the training of a classifier.

We use the K-means clustering algorithm for our experiments because it is simple and efficient. Assume that feature vectors of a particular data set are scattered in the vector space constituting clusters by virtue of similar features. The K-means algorithm tries to find vectors that are in the centers of these clusters. These special vectors are called centroids. Dependent on which centroid is closer to a vector in the vector space the vector will be assigned to the respective centroid's cluster. The

centroid $\vec{\mu}$ of a cluster ω is defined as follows:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

The overall goal of the algorithm is to minimize the distance between the centroids and the vectors in their respective clusters. To calculate the distances between the feature vectors and their centroids the Euclidean distance is used. The Euclidean distance between two vectors $\vec{x}$ and $\vec{y}$ is calculated by means of the following formula:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The K-means algorithm works as follows. After the number of target clusters is specified, the same number of seed points is placed randomly in the vector space. Then for all data points in the vector space the Euclidean distances to the seed points are calculated. Thus, the data points are assigned to the centroid which they have the shortest distance to constituting initial clusters. Next, for all data points belonging to a certain centroid a new centroid is calculated which represents a better approximation to the center of the respective cluster. Now the algorithm iterates. That is, it is repeated: The distances from the new centroids to the data points are calculated, data points are reassigned to other centroids if necessary and new centroids are calculated. The algorithm iterates until a certain number of iterations is completed or until there is no change in assignment of data points to the centroids compared to the previous iteration. Thus, in the end the algorithm yields information about the membership of each data point to a particular cluster.

In our experiments the normalized concatenated feature vectors are the data points in the vector space we want to find centroids for. Since it is known which adjective-noun phrase and thus which corresponding feature vector is literal or metaphorical, the assignment of the clustering algorithm can be evaluated in terms of precision, recall, F_1-measure and accuracy. The evaluation measures we use are explained in Subsection 4.1.3. To conduct our experiments we use the Weka[1] tool in the version 3.6.10. Weka provides various feature selection, classification and clustering algorithms as well as an implementation of the K-means algorithm. Moreover it has an option to visualize the clustering results which is useful for a detailed investigation of the influence of particular features. The experiments and their results are discussed in Chapter 4.

[1] http://www.cs.waikato.ac.nz/ml/weka/

3 Data and Pre-Processing

In Chapter 1 we formulated the following hypothesis:

While direct translation of a literal expression to another language will be mostly acceptable, the direct translation of a metaphoric expression will mostly fail.

The starting point for a verification of this hypothesis is to take literal and metaphorical expressions e. g. in English, to translate them into several other languages and to check how often these translated expressions occur in corpora for other languages. If the frequencies of literal expressions remain on a similar level across all languages but the frequencies of translated metaphorical expressions show significant differences the hypothesis can be seen as proven to be true.

The underlying method for a verification of the hypothesis is described in the preceding chapter. The current chapter gives extensive information about our data and its pre-processing.

To test our hypothesis three types of linguistic resources are required:

1. A test set of literal and metaphorical expressions that are labeled literal and metaphorical (Gold standard)
2. Bilingual dictionaries to translate the expressions of the test set into other languages.
3. Parallel and comparable corpora in several languages.

These resources and their origins are described in the following sections.

The selection of languages to perform experiments with is an important point which should be considered profoundly. First of all it seems to be important to choose languages that do not belong to the same family of languages. Closely related languages might exhibit similar usages of metaphors due to their common origin. By choosing languages from different families of languages we attempt to avoid such behavior. The second requirement is to avoid languages with close geographic vicinity. This would make the common use of metaphors less probable as well. One should despite be aware that borrowings can never be completely excluded. Even if two languages are not spoken geographically closely to each other. The

third requirement is that the author should have at least a basic knowledge of the used languages to be able to assess the translation results and to find irregularities which can arise for example from wrong encoding of the dictionaries or of the corpora. Languages that comply with these requirements are English, German, French, Spanish, Estonian and Russian. Although English and German do both belong to the family of Germanic languages they are still considered different enough for a comparison due to their geographical distance. French and Spanish both belong to the family of Romance languages but have a different history and can therefore be viewed as different enough to show discrepancies with regard to metaphors. Russian and Estonian, in turn, are geographically very close to each other but belong to completely different families of languages (Estonian: Finno-Ugric; Russian: Slavic) what makes them acceptable for our experiments.

3.1 Metaphor Test Set

The test set that is used for our experiments consists of 100 adjective-noun phrases. It is the same set that was used by Turney et al. (2011) and was kindly provided by Yair Neuman. Figure 3.1 shows an excerpt from this test set. The file is a CSV (Comma-separated values) file which uses commas to store "tabular data in plain text form"[1].

We will explain the content of each line by means of the first line. The first value is the literal or metaphorical expression we want to translate. The second value is the abstractness value as computed by the algorithm of Turney et al. (2011). The remaining values are decisions and ratings from five judges. Each judge first assigns a class to the expression and then a concreteness/abstractness rating. The number "1" stands for the literal class and "2" for the metaphorical class. The ratings of concreteness/abstractness can be higher. Judge No. 5 for example seems to consider the phrase *dark chocolate* to be very concrete while he does not rate the phrase *deep scepticism* as abstract as the other four judges do.

Since nouns in our dictionaries are only listed as singular forms, we modify the following three adjective-noun phrases of the metaphor data set for the sake of a better coverage:

- dark eyes → dark eye
- hard numbers → hard number
- warm feelings → warm feeling

To be able to evaluate the assignment of classes in the experiments later on, each adjective-noun phrase is assigned one label. This label is obtained by counting the class labels given by the five judges and then by taking the most

[1] http://en.wikipedia.org/wiki/Comma-separated_values

```
"dark chocolate",0.15145,1,1,1,1,1,1,1,1,1,2
"dark background",-0.0949,1,1,1,1,1,1,1,2,1,1
"dark suit",-0.08835,1,1,1,1,1,1,1,1,1,1
"dark figure",-0.08654,1,1,2,4,2,3,1,1,2,4
"deep sense",-0.40672,2,4,2,4,2,3,2,3,2,3
"deep respect",-0.30437,2,4,2,3,2,4,2,3,2,3
"deep red",0.31008,2,3,2,3,2,4,1,1,2,3
"deep distrust",-0.2611,2,4,2,4,2,3,2,3,2,3
"deep bowl",0.26965,1,1,1,1,1,1,1,1,1,1
"deep skepticism",-0.25039,2,4,2,4,2,4,2,4,2,3
```

Figure 3.1: Excerpt from the adjective-noun test set.

frequently assigned class label. Summing up the adjective-noun phrases with regard to these new labels it turns out that 44 phrases are literal and 56 metaphorical.

3.2 Dictionaries

To be able to translate adjective-noun phrases automatically it is necessary to dispose of bilingual dictionaries that are available locally. Dictionaries in a human readable format are preferred in order to be able to extend or manipulate them manually. Such dictionaries can then be read in by a self-developed tool that uses the dictionary entries to translate the adjective-noun phrases. This program is written in the Python programming language[2].

Several dictionaries available on the web as plain text files were evaluated in terms of how many words from the metaphor test set they were able to translate. Among them dictionaries from *dict.cc*[3], *BEOLINGUS*[4] (Chemnitz University of Technology), *Universal dictionary*, *Wiktionary*, *Omegawiki* (all three from *Dicts.info*[5]) and the Freelang project[6].

In order to translate from English to German and vice versa the BEOLINGUS dictionary offered satisfying results. Figure 3.2 shows an excerpt from the dictionary file. As can be observed from the excerpt the file has to undergo some modifications before it can be used for automatic translation. Our tool splits each line at the two colons to get the corresponding entries for German and English.

[2] http://www.python.org/
[3] http://www.dict.cc/
[4] http://dict.tu-chemnitz.de/, German-English dictionary file available at: http://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/
[5] http://www.dicts.info/uddl.php
[6] http://www.freelang.net/

In the next step gender and number markers (*m, f, pl*) are removed. In the final step the singular and plural forms are extracted. They are usually separated by the vertical bar and there is always the same number of vertical bars in each line.

```
Metallzuschnitt {m} :: metal blanc
Metamorphismus {m} :: metamorphism
Metanoia {f}; Änderung der Weltsicht :: shift of mind; change of mind; metanoia
Metapher {f}; übertragener Ausdruck | Metaphern {pl} :: metaphor | metaphors
Metaphorik {f} :: imagery
Metaphysik {f} :: metaphysics
```

Figure 3.2: Excerpt from the BEOLINGUS German-English dictionary file.

For the translations from English to French, Spanish, Estonian and Russian the dictionary from the Freelang project was the most comprehensive dictionary being able to translate almost all of the metaphors from the metaphor test set. Contrary to other dictionaries mentioned above, Freelang offers no option to download the dictionary files directly as plain text files. Instead, the user has to download a dictionary program which contains the dictionary files. These dictionary files look like binary files as demonstrated in Figure 3.3. After a short examination, though, it becomes obvious that the dictionary entries themselves are separated by binary NUL Bytes. Therefore a self-developed tool splits the file at the NUL Bytes, assigns two entries at a time to each other and writes them separated by a tab as a new line to a plain text file. Thus, we obtain a human readable dictionary in plain text format which is shown in Figure 3.4. This file can now be used in the same way as the plain text file obtained from BEOLINGUS. It is worth noting that the Freelang binary files do not all have the same encoding. By trying out different encodings we found out the encoding for the English-French and English-Spanish dictionaries to be Windows-1251, for the English-Estonian dictionary Latin-4 and for the English-Russian dictionary Windows-1251.

```
'em^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@neid(=
them)^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^
@^@^@^@^@^@^@^@^@^@'tis^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^
@^@^@^@^@^@^@^@see on (= it is)^@^@^@^@^@^@
```

Figure 3.3: Initial part of the Freelang English-Estonian dictionary (screenshot from the UNIX text viewer *less*).

```
'em          neid (= them)
'tis         see on (= it is)
```

Figure 3.4: Initial part of the Freelang English-Estonian dictionary after removal of NUL Bytes and after the alignment of two entries at a time.

3.3 Corpora

Corpora for several languages are the last resource which is required for our experiments. Similarly to the proper selection of languages the proper choice of corpora is an important point as well. As mentioned in Chapter 2 we use parallel corpora in addition to comparable corpora to make sure that differences in size or thematic disparities of the comparable corpora do not affect the results of the experiments in an unfavorable way. A parallel corpus is according to Koehn (2010) "a collection of text, paired with translations into another language". In contrast, comparable corpora are "corpora, where a series of monolingual corpora are collected for a range of languages, preferably using the same sampling frame and with similar balance and representativeness [...]" (McEnery, 2003). The origin of the corpora and their compilation process is described in the following subsections.

3.3.1 Parallel Corpora

Parallel corpora are mostly used in the domain of Statistical Machine Translation. Systems performing this kind of machine translation are usually trained on parallel corpora. During training algorithms search for word and phrase alignments in parallel corpora to automatically build lexica with word and phrase mappings from one language to another. We use some of the parallel corpora that have been made publicly available for the WMT translation task[7]. The WMT shared task is a venue for researchers who are developing systems for statistical machine translation. The shared task gives them the opportunity to translate a defined test set into other languages using their systems and to submit the results. These results are then evaluated automatically as well as by human judges.

We use three different parallel corpora for our experiments: The *Europarl* corpus[8], the *News Commentary* corpus [9] and the *Common Crawl* corpus [10]. The Europarl corpus is extracted from the proceedings of the debates of the European Parliament. The News Commentary comprises commentaries and articles about financial and political topics. The Common Crawl corpus consists mainly of various web pages from all across the web which have the same content but are available in different languages, for example web shops and homepages of institutions or organizations.

[7] http://www.statmt.org/wmt13/translation-task.html
[8] http://www.statmt.org/europarl/
[9] http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz
[10] http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz

Due to this composition the corpora are expected to be balanced. The corpora are tokenized and annotated with part-of-speech tags and word lemmas by Helmut Schmid's TreeTagger (Schmid, 1994). An overview of the sizes of the parallel corpora is given in Table 4.2 of Section 4.1.

3.3.2 Comparable Corpora

The Wikipedia Project "is a multilingual, web-based, free-content encyclopedia project based on an openly editable model."[11] Those properties make Wikipedia a good starting point for our experiments, since it almost perfectly meets the requirement on comparable corpora.

The part-of-speech annotated and lemmatized Wikipedia corpora for English and German from April 2011 have been kindly provided by André Blessing (Institute for Natural Language Processing, University of Stuttgart). The compilation of the Estonian and Russian Wikipedia corpora is described in the next subsection.

Processing of Estonian and Russian Wikipedia Corpora

Copies of Wikipedia articles databases can be obtained on the Wikipedia homepage[12]. Since they are encoded in XML[13] format as shown in Figure 3.5 they have to be preprocessed before they can be used for our experiments. That process is outlined below.

```
|colspan=2 align=center|&lt;div style=&quot;font-size:
90%;&quot;&gt;Vaata lähemalt selle artikli [[:{{NAMESPACE}} talk:
{{PAGENAME}}|aruteluleheküljelt]].&lt;/div&gt;
|}&lt;/div&gt;&lt;/center&gt;
*[[Ajalooline geograafia]] - [[Allikaõpetus]] - [[Antropoloogia]]
- [...]

==Etümoloogia==

Eestikeelne termin &quot;ajalugu&quot; on [[neologism]], [...]
```

Figure 3.5: Content of the Estonian Wikipedia articles dump (excerpt).

To extract single articles from the Russian and Estonian dumps we used the WP2TXT tool[14]. As Figure 3.6 shows, the output of WP2TXT is quite accurate but it still contains empty lines and some characters like the stars as list

[11] http://en.wikipedia.org/wiki/Wikipedia:About
[12] http://dumps.wikimedia.org/
[13] http://en.wikipedia.org/wiki/Xml
[14] http://wp2txt.rubyforge.org/

item markers that are useless for our purposes. Both have been removed by a self-developed tool. Additionally, long lists consisting of dates have been removed manually in order to accelerate the subsequent tagging process.

```
Asutuse sisemist töökorraldust reguleerivates dokumentides
määratakse asjaajamistoiminguid korraldavad (vastutavad)
struktuuriüksused ja ametnikud:

* Asjaajamise korraldamine (sh. asjaajamiskorra, dokumentide
loetelu koostamine)

* Dokumentide registreerimine ja ringluse korraldamine
```

Figure 3.6: Output of the WP2TXT tool applied to the Estonian Wikipedia dump.

As the next step the Wikipedia corpus is tokenized and annotated with part-of-speech tags and word lemmas by the TreeTagger. The TreeTagger option *-no-unknown* was added to ensure that TreeTagger outputs the word form rather than <*unknown*> for unknown lemmas. Figure 3.7 shows an excerpt from the Russian tagged Wikipedia Corpus.

To sum up we perform the following steps to process Estonian and Russian Wikipedia dumps. The resulting format is indicated in round brackets.

1. Download of Wikipedia-Dump (XML)
2. Executing WP2TXT (plain text)
3. Removal of unnecessary characters and empty lines (cleaned plain text)
4. Tokenization and part-of-speech tagging (tokenized one-word-per-line text annotated with lemmas and part-of-speech tags)

```
На          PR          на
нем          S          нем
видны          A            видный
несколько          NUM            несколько
кратеров          S          кратеров
размерами          S           размер
30-50          S          30-50
км          S          км
.          SENT           .
```

Figure 3.7: Extract from the part-of-speech tagged Russian Wikipedia corpus[15]

[15]The lemma of 'кратеров' is 'кратер', indeed. The TreeTagger did not know the lemma and therefore took the word form 'кратеров' as lemma (as indicated by the *-no-unknown* option).

Additional Corpora for Estonian and Russian

The Estonian and Russian Wikipedia corpora are much smaller than their English and German counterparts. Therefore, we add another corpora to compensate for this deficiency to a certain extent.

The Research Group of Computational Linguistics at the University of Tartu[16] provides corpora[17] that have been compiled from different sources like newspaper texts, fiction and scientific texts. Only a little subset of these corpora is morphologically annotated. Therefore we additionally used the Balanced Corpus[18] which is a subset of the Estonian Reference corpus that is currently under construction. These corpora were then annotated with part-of-speech tags and lemmas by the TreeTagger.

In addition to the Russian Wikipedia corpus we used a subset of a web corpus with 4.833.608 tokens and 47.643 lemmas available from the University of Leeds[19]. An overview of the sizes of the comparable corpora is given in Table 4.1 of Section 4.1.

3.3.3 Adjective-Noun Tuples

We do not really need the entire corpora in our experiments since we only extract features from particular adjective-noun phrases. Therefore we extract adjective-noun phrases from the corpora presented above by means of sequences of appropriate part-of-speech tags. Table 3.1 lists these sequences. In the end we get lists of all adjective-noun phrases occurring in the corpora. It is worth mentioning that in French and Spanish the adjective usually follows the noun. There are a few exceptions to this rule, though. For example, a little number of frequently used adjectives like *young* and *big* come usually before the noun. But since no adjectives from our metaphor test set (*deep*, *dark*, *hard*, *sweet*, *warm*) can be translated into one of these exceptional adjectives, this fact can be neglected. Therefore, for French and Spanish we extract only noun-adjective sequences. The numbers of extracted adjective-noun phrases are given in Table 4.3. Since we conduct separate experiments with parallel and comparable corpora and do not use comparable corpora for French and Spanish, no adjective-noun phrases have been extracted for them. Likewise, we did not extract Estonian and Russian adjective-noun phrases from parallel corpora because only data from comparable corpora is available for these two languages.

[16] http://www.cl.ut.ee/
[17] http://www.cl.ut.ee/korpused/
[18] http://www.cl.ut.ee/korpused/grammatikakorpus/
[19] http://corpus.leeds.ac.uk/mocky/

Language	Extracted sequences of part-of-speech tags
English	JJ* NN*
German	ADJ* NN
French	NOM ADJ
Spanish	NC ADJ
Estonian	A.* S.com*
Russian	A* S*

Table 3.1: Extracted sequences of part-of-speech tags. The wildcard * indicates that all part-of-speech tags were considered that started with the given prefix.

4 Experiments

The previous chapter depicts the process of obtaining the data needed for our experiments. The current chapter describes the conducted experiments anticipated in Chapter 2 using the obtained data. We start by exemplifying the experimental setup and present then the results of all conducted experiments accompanied by a discussion of these.

4.1 Experimental Setup

In the first part of this section we list the data used for our experiments. In the second part we give a detailed overview of the setup of the conducted experiments.

4.1.1 Data

As shown in preceding chapters we use a metaphor test set consisting of 100 English adjective-noun phrases of which 44 are literal and 56 metaphorical. These adjective-noun phrases are translated using bilingual dictionaries induced from online resources into German, French, Spanish, Estonian and Russian. Then we search for these translations in corpora of the respective language to extract their features. The experiments are conducted separately for the data originating from comparable and parallel corpora. We give the token counts for comparable corpora in Table 4.1. As can be observed from this table, the corpora differ greatly in size. Therefore we repeat all experiments conducted for the comparable corpora with parallel corpora for English, German, French and Spanish. This way we ensure that the corpora for the different languages have the same content. Their numbers of tokens are given in Table 4.2. Since we only need adjective-noun phrases for the feature extraction of the translations, we extract these phrases from the corpora as described in the preceding chapter. The counts of adjective-noun phrases for each language and corpus type are given in Table 4.3.

4.1.2 Conducted Experiments

As depicted in section 2.2 we make use of 7 types of feature vectors to carry out our experiments on:

1. "Best-Translation": One translation selected out of all translations of an English phrase.

Language	Corpus	Number of tokens	
		per corpus	per language
English	Wikipedia	935 038 310	935 038 310
German	Wikipedia	432 131 420	432 131 420
Russian	Wikipedia	23 826 335	28 659 943
	Mocky	4 833 608	
Estonian	Wikipedia	21 913 998	40 264 529
	Morphol. disamb. corpus	624 582	
	Balanced corpus	17 725 949	

Table 4.1: Number of tokens of comparable corpora

Language	Number of tokens			
	Europarl	News Commentary	Common Crawl	per language
English	50 263 003	3 949 846	70 727 227	124 940 076
German	44 613 020	4 054 215	47 045 739	95 712 974
French	52 525 000	4 086 635	76 688 347	133 299 982
Spanish	51 622 215	4 595 283	43 514 857	99 732 355

Table 4.2: Number of tokens of parallel corpora

2. “Mean”: Mean of feature values of all translations.
3. “Mean (w/o 0)”: “As Mean”, zeros excluded.
4. “Median”: Median of feature values of all translations.
5. “Median (w/o 0)”: As “Median”, zeros excluded.
6. “Std”: Standard deviation of feature values of all translations.
7. “Std (w/o 0)”: As “Std”, zeros excluded.

As mentioned before, all experiments are separately conducted on data originating from comparable corpora for English, German, Estonian and Russian as well as on parallel corpora compiled from English, German, French and Spanish data.

4.1.3 Baselines and Evaluation Measures

We define two baselines based on cluster distribution assumptions. If our hypothesis is wrong then the feature value differences should be such insignificant that the clustering algorithm is not able not separate them into two clusters. Instead it would put all vectors into one cluster while the other cluster would remain empty. Our test set consists of 44 literal and 56 non-literal adjective-noun phrases that are to be clustered. The metaphorical class is the major class. Thus, our first baseline are two clusters with the following distribution: one metaphorical cluster that contains all vectors and one empty literal cluster. We refer to this baseline as

Language	Abbr.	Comparable corpora	Parallel corpora	Total
English	EN	39 906 011	6 979 464	46 885 475
German	DE	24 139 616	6 075 299	30 214 915
French	FR	–	5 586 944	5 586 944
Spanish	ES	–	3 887 587	3 887 587
Estonian	EE	1 685 538	–	1 685 538
Russian	RU	2 597 832	–	2 597 832

Table 4.3: Numbers of extracted adjective-noun phrases

"Majority".

The second baseline is based on the assumption that the vectors are randomly assigned to the two clusters because the features do not show any significant difference between literal vectors and metaphorical vectors. In this case we assume two clusters with evenly distributed feature vectors where each cluster contains 22 literal and 28 metaphorical vectors. This baseline is referred to as "Random".

We show the accuracy for both classes and precision, recall and the F_1-measure for the literal and metaphorical classes separately and compare them to our baselines. Following Birke and Sarkar (2006) we define literal precision and literal recall as follows:

$$literal\_precision = \frac{correct\ literals\ in\ literal\ cluster}{size\ of\ literal\ cluster}$$

$$literal\_recall = \frac{correct\ literals\ in\ literal\ cluster}{total\ literals}$$

Metaphorical precision and recall is defined analogously:

$$metaphorical\_precision = \frac{correct\ metaphors\ in\ metaphorical\ cluster}{size\ of\ metaphorical\ cluster}$$

$$metaphorical\_recall = \frac{correct\ metaphors\ in\ metaphorical\ cluster}{total\ metaphors}$$

The F_1-measure combines precision and recall into an overall measure. It is defined as follows:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

Therefore, we define the F_1-measure for the literal cluster as:

$$literal\_F_1 = \frac{2 \times literal\_precision \times literal\_recall}{literal\_precision + literal\_recall}$$

Again, the F_1-measure for the metaphorical cluster is defined analogously:

$$metaphorical\_F_1 = \frac{2 \times metaphorical\_precision \times metaphorical\_recall}{metaphorical\_precision + metaphorical\_recall}$$

The accuracy is defined as:

$$accuracy = \frac{correct\ literals\ in\ literal\ cluster\ +\ correct\ metaphors\ in\ metaphorical\ cluster}{number\ of\ instances\ in\ both\ clusters}$$

4.2 Results

This section presents the results of our experiments. The experimental setup as well as the evaluation measures are explained in the preceding section. In the tables below "L" stands for the literal cluster, "M" for the metaphorical cluster and "Avg." for the average of both results.

4.2.1 Frequency of Translations

Our initially formulated hypothesis assumed that literal adjective-noun phrases would have a similar frequency in other languages while metaphorical phrases would not appear at all or would have a much lower frequency. Thus, in the first experiment we verify to what extent the frequencies of translations of adjective-noun phrases can prove this hypothesis true. Therefore, we conduct clustering experiments for data originating from the comparable corpora and from parallel corpora where only the four *freq* features of the concatenated feature vectors are used. The results of the clustering are given in Table 4.4 for comparable corpora and in Table 4.5 for parallel corpora.

As can be seen from Table 4.4 for the comparable corpora, the averages for precision, recall and the accuracy are always higher than the two baselines. It can also be noted that using the median to compute the averages increases precision and accuracy. The recall of the literal cluster is quite low while the recall of the

	Precision			**Recall**			**F_1-Measure**			**Accuracy**
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.636	0.584	0.610	0.159	0.929	0.544	0.255	0.717	0.486	0.590
Mean	0.556	0.571	0.563	0.114	0.929	0.521	0.189	0.707	0.448	0.570
Mean (w/o 0)	0.563	0.583	0.573	0.205	0.875	0.540	0.300	0.700	0.500	0.580
Median	0.800	0.579	0.689	0.091	0.982	0.537	0.163	0.728	0.446	0.590
Median (w/o 0)	0.600	0.588	0.594	0.205	0.893	0.549	0.305	0.709	0.507	0.590
Std	0.615	0.586	0.601	0.182	0.911	0.546	0.281	0.713	0.497	0.590
Std (w/o 0)	0.545	0.573	0.559	0.136	0.911	0.524	0.218	0.703	0.461	0.570

Table 4.4: Comparable corpora: Results for the *freq* feature (frequencies of adjective-noun phrase translations)

	Precision			**Recall**			**F_1-Measure**			**Accuracy**
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.500	0.571	0.536	0.182	0.857	0.519	0.267	0.686	0.476	0.560
Mean	0.667	0.591	0.629	0.182	0.929	0.555	0.286	0.722	0.504	0.600
Mean (w/o 0)	0.529	0.578	0.554	0.205	0.857	0.531	0.295	0.691	0.493	0.570
Median	0.583	0.580	0.581	0.159	0.911	0.535	0.250	0.708	0.479	0.580
Median (w/o 0)	0.533	0.576	0.555	0.182	0.875	0.528	0.271	0.695	0.483	0.570
Std	0.636	0.584	0.610	0.159	0.929	0.544	0.255	0.717	0.486	0.590
Std (w/o 0)	0.800	0.600	0.700	0.182	0.964	0.573	0.296	0.740	0.518	0.620

Table 4.5: Parallel corpora: Results for the *freq* feature (frequencies of adjective-noun phrase translations)

metaphorical cluster is similar to the first baseline which assumes that all phrases are assigned to the larger metaphorical cluster. Therefore it can be assumed that the higher number of metaphorical phrases causes a bias towards the metaphorical cluster.

The results for the parallel corpora show a similar picture. But almost all values for the Best-Translation experiment are worse. Obviously the heuristic for the computation of the best translation does not work that well for parallel corpora. The results for the *Mean* and the *Std (w/o 0)* experiments exhibit a clear improvement compared to the comparable corpora whereas the other experiments yield almost equal or worse scores. Therefore it can be assumed that *Mean* and *Std (w/o 0)* benefit from the better balanced underlying data. *Std (w/o 0)* benefits in particular from the removal of feature values which equal to 0.0. To sum up we can state that translating English adjective-noun phrases merely slightly helps to separate literal and metaphorical phrases.

4.2.2 Effects of the Versatility of the Noun

As mentioned in Chapter 2 we also include two features to verify whether words appearing in many contexts tend to be used metaphorically more often. The feature *ADJ-nn* captures the amount of different adjectives the noun can occur with while *adj-NN* is the number of different nouns the adjective can occur with. We carry out two types of experiments[1]

1. Both features (*ADJ-nn* and *adj-NN*) are used for clustering. Other features such as *freq* are not included.
2. Only the *ADJ-nn* feature is used for clustering. Again, other features such as *freq* are not included.

Tables 4.6 and 4.7 show the results for experiments with comparable corpora while Tables 4.8 and 4.9 present the results for the parallel corpora. We discuss the results for comparable corpora first. As can be observed from Table 4.6 the results are very close to the second baseline with the exception of the Best-Translation experiment. The Best-Translation experiment scores almost always the highest values here. The reason for this is, that no averages over the *adj-NN* feature values are computed as in the *Mean*, *Median* and *Std* experiments. Obviously the clustering algorithm is confused by data where averages over *adj-NN* values are computed which are often the same due to the low number of adjectives in the

[1] We do not carry out experiments, where merely the *adj-NN* feature is used due to the structure of our metaphor test set. Since it consists of 100 adjective-noun phrases which all contain different nouns but only 5 different adjectives (*dark*, *deep*, *hard*, *sweet* and *warm*), the feature would be too general and could not provide sufficient distinctive information to separate literal from metaphorical phrases.

metaphor test set. Also the differences in size of the underlying corpora might play a role here. If, for example, less translations can be found in smaller corpora, the averages in the experiments based an averaged data are computed on less elements which can introduce an undesirable bias between the features for the single languages.

If we compare the experiments where only the *ADJ-nn* feature is considered (Table 4.7) to the results from Table 4.6, we see an improvement of both the literal and the metaphorical cluster in terms of precision and recall with the exception of the recall of the metaphorical cluster. This suggests that the amount of different adjectives a noun can occur with is indeed a promising feature for literal adjective-noun phrases.

The tendency of a slight improvement of the precision of the literal cluster when using only the *ADJ-nn* feature can also be observed for the experiments with the parallel corpora (Tables 4.8 and 4.9). But in contrast to the experiments with the comparable corpora now the recall of the metaphorical cluster benefits from the only use of the *ADJ-nn* feature (Table 4.9) instead of the recall of the literal cluster. The single use of the *ADJ-nn* feature instead of a combination of the *ADJ-nn* and *adj-NN* features improves the overall accuracy and F_1-Measure. The *Mean* and *Std* settings score thereby even above the first baseline.

To sum up, we can state that the mere consideration of the number of different adjectives a noun can occur with helps to separate literal from metaphorical phrases. But since our experimental setups not only differ in size of the used corpora but also in the used languages our results cannot give a clear statement about the exact cause of the improvement. Therefore, the influence of this feature requires further research.

4.2.3 Influence of the Association Between Adjective and Noun

Our final experiments conducted on data collected from translations into different languages examine the influence of two association measures, namely the Pointwise Mutual Information (PMI) and the chi-square test (χ^2). As mentioned in Chapter 2 these features capture the information about the association of two words.

Similar to other experiments described above we conduct separate experiments for data from comparable corpora and for data from parallel corpora. First the individual performances of the PMI feature and of the χ^2 feature are examined. Then the performance of both features in combination is investigated.

The performance of the PMI feature for the experiments with comparable corpora is given in Table 4.10 and for the experiments with parallel corpora in Table 4.11. We can observe that the *Best-Translation* setting performs much better on data

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.512	0.614	0.563	0.500	0.625	0.563	0.506	0.619	0.563	0.570
Mean	0.444	0.565	0.505	0.545	0.464	0.505	0.490	0.510	0.500	0.500
Mean (w/o 0)	0.455	0.578	0.516	0.568	0.464	0.516	0.505	0.515	0.510	0.510
Median	0.455	0.578	0.516	0.568	0.464	0.516	0.505	0.515	0.510	0.510
Median (w/o 0)	0.455	0.578	0.516	0.568	0.464	0.516	0.505	0.515	0.510	0.510
Std	0.458	0.577	0.518	0.500	0.536	0.518	0.478	0.556	0.517	0.520
Std (w/o 0)	0.447	0.565	0.506	0.386	0.625	0.506	0.415	0.593	0.504	0.520

Table 4.6: Comparable corpora: Results for the *ADJ-nn* and *adj-NN* features.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.472	0.643	0.558	0.773	0.321	0.547	0.586	0.429	0.507	0.520
Mean	0.478	0.645	0.562	0.750	0.357	0.554	0.584	0.460	0.522	0.530
Mean (w/o 0)	0.472	0.643	0.558	0.773	0.321	0.547	0.586	0.429	0.507	0.520
Median	0.465	0.621	0.543	0.750	0.321	0.536	0.574	0.424	0.499	0.510
Median (w/o 0)	0.465	0.621	0.543	0.750	0.321	0.536	0.574	0.424	0.499	0.510
Std	0.333	0.541	0.437	0.114	0.821	0.468	0.169	0.652	0.411	0.510
Std (w/o 0)	0.481	0.714	0.598	0.864	0.268	0.566	0.618	0.390	0.504	0.530

Table 4.7: Comparable corpora: Results for the *ADJ-nn* feature.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.494	0.762	0.628	0.886	0.286	0.586	0.634	0.416	0.525	0.550
Mean	0.486	0.667	0.576	0.773	0.357	0.565	0.596	0.465	0.531	0.540
Mean (w/o 0)	0.494	0.762	0.628	0.886	0.286	0.586	0.634	0.416	0.525	0.550
Median	0.475	0.700	0.588	0.864	0.250	0.557	0.613	0.368	0.491	0.520
Median (w/o 0)	0.475	0.700	0.588	0.864	0.250	0.557	0.613	0.368	0.491	0.520
Std	0.527	0.808	0.667	0.886	0.375	0.631	0.661	0.512	0.587	0.600
Std (w/o 0)	0.469	0.684	0.577	0.864	0.232	0.548	0.608	0.347	0.477	0.510

Table 4.8: Parallel corpora: Results for the *ADJ-nn* and *adj-NN* features.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.493	0.704	0.598	0.818	0.339	0.579	0.615	0.458	0.537	0.550
Mean	0.538	0.743	0.641	0.795	0.464	0.630	0.642	0.571	0.607	0.610
Mean (w/o 0)	0.530	0.735	0.633	0.795	0.446	0.621	0.636	0.556	0.596	0.600
Median	0.500	0.676	0.588	0.750	0.411	0.580	0.600	0.511	0.556	0.560
Median (w/o 0)	0.500	0.714	0.607	0.818	0.357	0.588	0.621	0.476	0.548	0.560
Std	0.527	0.808	0.667	0.886	0.375	0.631	0.661	0.512	0.587	0.600
Std (w/o 0)	0.521	0.759	0.640	0.841	0.393	0.617	0.643	0.518	0.581	0.590

Table 4.9: Parallel corpora: Results for the *ADJ-nn* feature.

from parallel corpora. The cause for this improvement seems to be the increase in precision and recall of the literal cluster and in precision of the metaphorical cluster. They make up for the decrease in recall of the metaphorical cluster and thus lead to a better performance. With regard to the experiments based on averaged data we can state that *Mean* and *Median* perform much better on data where zeros have been excluded from the computation. On data from parallel corpora, in the *Median (w/o 0)* experiment, we achieve the highest accuracy of all our experiments (0.73). The reason for this considerable improvement can be exemplified by means of the feature values of the German PMI feature in the Median setting. In the data for the experiment with included zeros only 2 out of 100 feature vectors have non-zero values for the PMI feature. In contrast, in the data for the experiment with excluded zeros 79 feature vectors out of 100 have non-zero values. The greater number of non-zero feature values provides more helpful data for the clustering algorithm and therefore improves the clustering result. *Mean* and *Median* experiments on comparable corpora cannot achieve the level of parallel corpora but are still mostly above the baseline. However, the *Std* experiments on both types of corpora show a drop in performance between the *Std* and *Std (w/o 0)* experiment. We will be discussing this peculiarity at the end of this section.

Contrary to the results for the PMI feature the individual performance of the χ^2 feature shows very mixed results. This can be observed in Tables 4.12 and 4.13. The *Best-Translation* setting merely yields a result which is above the baseline in the experiment based on the data from comparable corpora. The experiments based an averaged feature values (*Mean*, *Median*, *Std*) show an ambivalent picture. While the distribution of results collected on data from the comparable corpora is consistent with the distribution of results of the PMI experiments presented above the results of the *Mean* and *Median* settings based an data from parallel corpora do not fit anymore. So regarding the experiments based an data from comparable corpora the settings with excluded zeros show a better performance than the settings where the zeros are included. In contrast, in the experiments using data from parallel corpora the experiments where zeros have been excluded show a decline in performance. These results seem incomprehensible. In particular, since the exclusion of zeros yields more non-zero values analogously to the PMI feature. The explanation for the performance decline could be instead found in the overall distribution of results found in Table 4.13. For example, the results of the recall of the literal and metaphorical clusters are very similar to the "Majority" baseline while the result for the precision (except the *Median* setting) is similar to the "Random" baseline. This suggests that the χ^2 feature values collected on data from parallel corpora do not really capture a distinction between literal and metaphorical phrases.

Finally, we conduct experiments where the PMI and the χ^2 features are used in combination. Table 4.14 shows the results for the comparable corpora. Interestingly, here we achieve higher results than by considering the PMI or the

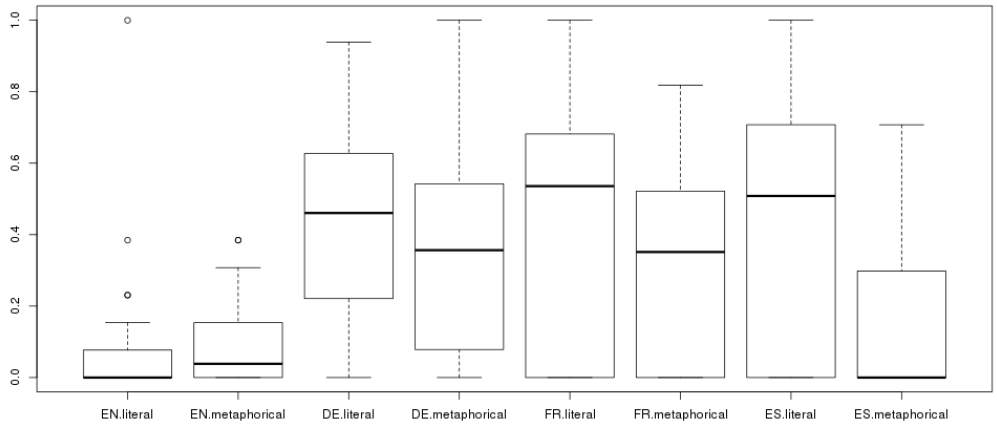

Figure 4.1: Boxplot for the PMI feature value distribution in the *Std* setting.

χ^2 feature alone. The same observation can be made for almost all experiments based on the parallel corpora. Their results are given in Table 4.15. Moreover, the results for the parallel corpora show a further improvement (except the *Best-Translation* and the *Std (w/o 0)* settings) compared to the experiments based on comparable corpora. In the *Mean* setting we achieve an accuracy of 0.72 and an average F_1-measure of 0.708 what is not only far beyond the baseline but also the second highest accuracy (0.72) and average F_1-measure (0.708) that we achieve in all our experiments. Therefore it can be concluded that the combination of the PMI and the χ^2 features benefits in particular from the balanced corpora.

The large drop in performance between the *Std* and the *Std (w/o 0)* experiments based on the parallel corpora using the PMI feature or the combination of the PMI and the χ^2 features needs further investigation. Figure 4.1 shows a boxplot diagram visualizing the distribution of feature values of the PMI feature for each language divided according to the literal and metaphorical class. Figure 4.2 shows the boxplot diagram for the *Std (w/o 0)* experiment.

It is noticeable that the range of feature values of the Spanish metaphorical vectors has substantially diminished in the *Std (w/o 0)* setting compared to the *Std* setting. A manual analysis of the PMI feature values confirmed that the Spanish PMI feature in the vector used for the *Std (w/o 0)* experiment contains much more 0.0 values than the vector used for the *Std* experiment. To prove that the Spanish feature is the main cause for the performance drop we conducted two further experiments where merely the English and Spanish PMI and χ^2 features were used. The results of these experiments are given in Table 4.16. As can be observed from the table, the mere use of English and Spanish PMI and χ^2 features causes a performance drop of 0.19 in accuracy and 0.274 in the average F_1-measure between the *Std* and *Std*

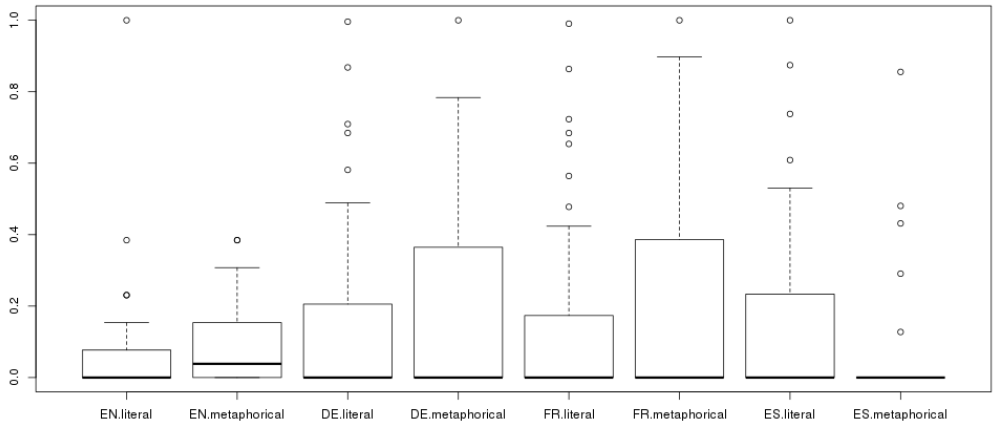

Figure 4.2: Boxplot for the PMI feature value distribution in the *Std (w/o 0)* setting.

without zeros setting. This shows that the feature values of the Spanish PMI feature are negatively affected when calculating their standard deviation which leads to an overall worse performance.

4.2.4 Influence of Translation Features

The results presented in the preceding section suggest that association features are the most reliable features for separating literal from metaphorical adjective-noun phrases. Therefore the question arises whether merely the use of the PMI feature computed for the original English phrases could suffice in order to reliably separate literal and metaphorical phrases. Thus we conduct an additional experiment to verify to which extent the PMI feature of the original English phrases can substitute the usage of feature values computed from translations into other languages.

In Table 4.17 the results for the experiments merely using the PMI feature for English with the parallel corpora are given. As can be observed, almost all settings outperform the baselines in terms of the average F_1-measure and accuracy. Only the *Std* experiments score on the baseline level. Interestingly, the *Best-Translation* setting scores better than the similar setting shown in Table 4.11. But all experiments based on averaged feature values (except *Std (w/o 0)*) show a decrease in performance compared to the similar experiments shown in Table 4.11 which uses the PMI features extracted from the English phrases as well as from their translations. Therefore we can conclude that features capturing the association strength between two words extracted from the translations of the original phrase can significantly improve the separation of literal and metaphorical adjective-noun phrases in English.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.333	0.541	0.437	0.114	0.821	0.468	0.169	0.652	0.411	0.510
Mean	0.464	0.688	0.576	0.886	0.196	0.541	0.609	0.306	0.457	0.500
Mean (w/o 0)	0.622	0.709	0.666	0.636	0.696	0.666	0.629	0.703	0.666	0.670
Median	0.552	0.606	0.579	0.364	0.768	0.566	0.438	0.677	0.558	0.590
Median (w/o 0)	0.596	0.698	0.647	0.636	0.661	0.649	0.615	0.679	0.647	0.650
Std	0.558	0.688	0.623	0.659	0.589	0.624	0.604	0.635	0.619	0.620
Std (w/o 0)	0.537	0.627	0.582	0.500	0.661	0.580	0.518	0.643	0.581	0.590

Table 4.10: Comparable corpora: Results for the *PMI* feature.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.524	0.703	0.613	0.750	0.464	0.607	0.617	0.559	0.588	0.590
Mean	0.667	0.644	0.655	0.409	0.839	0.624	0.507	0.729	0.618	0.650
Mean (w/o 0)	0.767	0.700	0.733	0.523	0.875	0.699	0.622	0.778	0.700	0.720
Median	0.667	0.644	0.655	0.409	0.839	0.624	0.507	0.729	0.618	0.650
Median (w/o 0)	0.743	0.723	0.733	0.591	0.839	0.715	0.658	0.777	0.718	0.730
Std	0.633	0.745	0.689	0.705	0.679	0.692	0.667	0.710	0.688	0.690
Std (w/o 0)	0.467	0.640	0.553	0.795	0.286	0.541	0.588	0.395	0.492	0.510

Table 4.11: Parallel corpora: Results for the *PMI* feature.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.857	0.591	0.724	0.136	0.982	0.559	0.235	0.738	0.487	0.610
Mean	1.000	0.577	0.789	0.068	1.000	0.534	0.128	0.732	0.430	0.590
Mean (w/o 0)	1.000	0.583	0.792	0.091	1.000	0.545	0.167	0.737	0.452	0.600
Median	1.000	0.577	0.789	0.068	1.000	0.534	0.128	0.732	0.430	0.590
Median (w/o 0)	1.000	0.589	0.795	0.114	1.000	0.557	0.204	0.742	0.473	0.610
Std	0.667	0.574	0.621	0.091	0.964	0.528	0.160	0.720	0.440	0.580
Std (w/o 0)	0.500	0.563	0.531	0.045	0.964	0.505	0.083	0.711	0.397	0.560

Table 4.12: Comparable corpora: Results for the χ^2 feature.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.500	0.570	0.535	0.159	0.875	0.517	0.241	0.690	0.466	0.560
Mean	0.417	0.557	0.487	0.114	0.875	0.494	0.179	0.681	0.430	0.540
Mean (w/o 0)	0.412	0.554	0.483	0.159	0.821	0.490	0.230	0.662	0.446	0.530
Median	0.583	0.580	0.581	0.159	0.911	0.535	0.250	0.708	0.479	0.580
Median (w/o 0)	0.417	0.557	0.487	0.114	0.875	0.494	0.179	0.681	0.430	0.540
Std	0.357	0.547	0.452	0.114	0.839	0.476	0.172	0.662	0.417	0.520
Std (w/o 0)	0.273	0.539	0.406	0.068	0.857	0.463	0.109	0.662	0.386	0.510

Table 4.13: Parallel corpora: Results for the χ^2 feature.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.574	0.717	0.646	0.705	0.589	0.647	0.633	0.647	0.640	0.640
Mean	0.464	0.688	0.576	0.886	0.196	0.541	0.609	0.306	0.457	0.500
Mean (w/o 0)	0.622	0.709	0.666	0.636	0.696	0.666	0.629	0.703	0.666	0.670
Median	0.600	0.613	0.607	0.341	0.821	0.581	0.435	0.702	0.569	0.610
Median (w/o 0)	0.596	0.698	0.647	0.636	0.661	0.649	0.615	0.679	0.647	0.650
Std	0.577	0.708	0.643	0.682	0.607	0.644	0.625	0.654	0.639	0.640
Std (w/o 0)	0.548	0.638	0.593	0.523	0.661	0.592	0.535	0.649	0.592	0.600

Table 4.14: Comparable corpora: Results for the *PMI* and χ^2 features.

	Precision			Recall			F_1-Measure			Accuracy
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.532	0.711	0.621	0.750	0.482	0.616	0.623	0.574	0.599	0.600
Mean	0.679	0.653	0.666	0.432	0.839	0.636	0.528	0.734	0.631	0.660
Mean (w/o 0)	0.722	0.719	0.720	0.591	0.821	0.706	0.650	0.767	0.708	0.720
Median	0.679	0.653	0.666	0.432	0.839	0.636	0.528	0.734	0.631	0.660
Median (w/o 0)	0.735	0.712	0.724	0.568	0.839	0.704	0.641	0.770	0.706	0.720
Std	0.608	0.735	0.671	0.705	0.643	0.674	0.653	0.686	0.669	0.670
Std (w/o 0)	0.474	0.667	0.570	0.818	0.286	0.552	0.600	0.400	0.500	0.520

Table 4.15: Parallel corpora: Results for the *PMI* and χ^2 features.

	Precision			**Recall**			F_1**-Measure**			**Accuracy**
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Std	0.742	0.696	0.719	0.523	0.857	0.690	0.613	0.768	0.691	0.710
Std (w/o 0)	0.357	0.547	0.452	0.114	0.839	0.476	0.172	0.662	0.417	0.520

Table 4.16: Parallel corpora: Results for the English and Spanish *PMI* and χ^2 features in the *Std* setting.

	Precision			**Recall**			**F_1-Measure**			**Accuracy**
	L	M	Avg.	L	M	Avg.	L	M	Avg.	
Majority	0.0	0.56	0.28	0.0	1.0	0.5	0.0	0.718	0.359	0.56
Random	0.44	0.56	0.5	0.5	0.5	0.5	0.468	0.528	0.498	0.5
Best-Translation	0.556	0.757	0.656	0.795	0.500	0.648	0.654	0.602	0.628	0.630
Mean	0.531	0.722	0.627	0.773	0.464	0.619	0.630	0.565	0.597	0.600
Mean (w/o 0)	0.537	0.758	0.647	0.818	0.446	0.632	0.649	0.562	0.605	0.610
Median	0.592	0.706	0.649	0.659	0.643	0.651	0.624	0.673	0.648	0.650
Median (w/o 0)	0.565	0.763	0.664	0.795	0.518	0.657	0.660	0.617	0.639	0.640
Std	0.375	0.548	0.461	0.136	0.821	0.479	0.200	0.657	0.429	0.520
Std (w/o 0)	0.385	0.552	0.468	0.114	0.857	0.485	0.175	0.671	0.423	0.530

Table 4.17: Parallel corpora: Results for the *PMI* feature from English phrases.

5 Related Work

Several different approaches to metaphor identification, or metaphor recognition, have been undertaken so far. They can be subclassified into knowledge-based approaches and statistical approaches. Knowledge-based approaches usually use a large, mostly hand-crafted knowledge base which contains rules for possible knowledge transfer from one domain into another domain. Usually, knowledge-based approaches work for a limited number of domains due to the large effort which is necessary to compile the knowledge bases. Statistical approaches, in turn, rely on machine learning techniques to "learn" from annotated resources about knowledge transfer taking place to create metaphorical expressions. Often clustering methods are applied as well to separate literal and metaphorical expressions. In the following, we present an overview of several relevant approaches to metaphor identification and discuss them.

Dolan (1995) describes a system for metaphor interpretation which exploits a lexical knowledge base derived from a machine-readable dictionary of English. The used dictionary contains not only word entries but also definition strings for each word. A semantic analysis of these definitions yields a lexical knowledge base consisting of semantic relations between the headwords and the words of their definition texts. Those relations contain, among others, relations like *Hypernym of* or *Part of*. Additionally, a disambiguation of the word senses takes place. Then, sets of typical objects for different senses of a verb are identified by means of the semantic relations. For example, the verb *plant* has two senses. The set of typical objects for the first sense consists of botanical words like *seed*, *grove* or *plantation* which are all linked to the noun *plant*. The set of typical objects for the second sense of *plant* consists of the nouns *belief* and *idea*. The intuition that these two sets are metaphorically connected to each other is proven by calculating paths between them. Several paths connect e. g. the word *seed* from the first set and the word *idea* from the second set. The noun *germ* is found out to be a frequent intersection on paths which connect *seed* and *idea*. The author argues that such words may reflect "pervasive metaphorical associations" between two concepts.

When Dolan's system encounters novel instances of metaphor it applies the same method it uses to discover metaphorical connections between e. g. verbs depicted above. For example, in the sentence *The idea flourished* the paths connecting the verb *flourish* and the noun *idea* are explored. Unfortunately, the author presents only selected examples and does neither conduct a broad evaluation of his system

nor does he give results for the interpretation of a larger set of metaphorical expressions.

Shutova et al. (2012) present a minimally supervised metaphor identification and interpretation system which according to their description "discovers literal meanings of metaphorical expressions in text and produces their literal paraphrases". As a starting point they compile a seed set consisting of 62 verb–object and verb–subject phrases like "throw remark", "tension mounted" or "example illustrates". Then they perform verb and noun clustering. For every seed expression a verb cluster is chosen representing the source concept and a noun cluster representing the set of possible target concepts. By linking these clusters metaphorical associations are formed.

In the metaphor identification task the text is first parsed to discover verb–object and verb–subject phrases. The system then checks whether the phrase terms occur in the previously linked clusters. If they do they are marked as metaphorical. This metaphor identification module is then evaluated in different modes. First, 38 randomly selected phrases are annotated by the system as well as by five human annotators. The human annotations are considered the gold standard. Comparing the system annotations to this gold standard the system achieves a precision of 0.79. When comparing the system annotations to every annotator separately and then calculating the average precision the system still achieves a precision of 0.74. Another evaluation is performed on 200 sample phrases which are annotated by the system and by one of the authors. In this experiment the system achieves a precision of 0.76. So the evaluation is either performed on a little test set or on a larger test set that is annotated only by one annotator which might only be reliable to a limited extent. Therefore it must be noted that no broad evaluation on a large data set takes place.

To find out proper substitutes for metaphoric phrases, the system of Shutova et al. (2012) runs through a three-step process. Since phrases are considered where the verb is used metaphorically only a lexical replacement of the verb takes place. First the most probable verbs in the given context are generated and ranked after their likelihood in a corpus. Secondly, common features between the metaphorically used verb and the verbs generated in the previous step are identified using the WordNet hierarchy. In this step unrelated paraphrases are filtered out. Finally, the paraphrases are ranked due to their selection preferences. On a test set with 62 subject–verb and verb–direct object constructions the metaphor interpretation module achieves a precision of 0.81, on average.

The modules for metaphor identification and metaphor interpretation are finally evaluated as an integrated system in terms of accuracy. As for the evaluation of the metaphor identification module this integrated system is also evaluated in two modes. Firstly, three annotators are presented with 35 sentences containing

metaphors. Their annotations are again considered the gold standard. The comparison of this gold standard to the integrated system's annotations yield an accuracy of 0.71. In a second evaluation setting a sample of 600 sentences is annotated by one judge. Here the system scores an accuracy of 0.67. A detailed error analysis conducted by the authors shows that the identification module performs with an accuracy of 0.72 while the interpretation module scores an accuracy of 0.68.

The approach of Shutova et al. (2012) is an interesting statistical endeavor to metaphor identification and interpretation which does not depend on hand-coded rules and operates open-domain. However, it needs a manually compiled seed set and a large amount of other data resources. Unfortunately, the way in which the evaluation is performed is not very well applicable to other systems of this kind due to the various sizes of the test sets.

Martin (1992) presents a knowledge-based approach to handle conventional metaphors like "How can I enter Lisp?". Conventional metaphors allow to express computer processes in terms of real world processes or entities. In the example above the program "Lisp" is viewed as an enclosure that can be entered to activate it. Martin's approach follows the principle "that the interpretation of metaphoric language should proceed through the direct application of specific knowledge about the metaphors in the language". He calls it the *Metaphoric Knowledge* approach and implements it in a system named MIDAS (*Metaphor Interpretation, Denotation, and Acquisition System*). MIDAS is aimed to represent knowledge about conventional metaphors, to apply this knowledge to interpret metaphors and and to learn new metaphors. To test the system, it is integrated into UNIX Consultant, a natural language consultant system that is intended to provide help to users that are new to the UNIX operating system.

The individual metaphors in MIDAS are represented as source concepts, target concepts and sets of associations between them. This knowledge is formalized by KODIAK, an extended semantic network language. When carrying out metaphor interpretation, the knowledge base is searched for appropriate interpretations which match the concepts and do not violate the semantic constraints. To a certain extent MIDAS is able to learn new metaphors. When the system encounters an unknown metaphor it tries to find the most similar interpretation whose target concept can then be extended to offer an explanation for the new metaphor.

Furthermore MIDAS complies with two constraints that follow from possible interpretations of results from psycholinguistic research. The first result states that the time needed to process metaphorical language does not differ significantly from the time needed to interpret literal language. This is known as the total-time-constraint. That leads to the constraint that the mechanisms used to process non-literal language should be basically the same as those for processing literal language. MIDAS meets this requirement by viewing non-literal processes

as conventional expressions and not as derivations of their literal counterparts. In this way metaphorical and non-metaphorical expressions can be interpreted using the same mechanisms. The second constraint draws on the observation that metaphorical interpretations are also feasible in contexts that already show a well-formed literal interpretation. For example, the phrase "McEnroe killed Connors" can be literally interpreted in the way that McEnroe did something that caused Connors' demise while the non-literal interpretation would mean that McEnroe defeated Connors in a competition. Therefore MIDAS retrieves all available interpretation in a given context.

Due to the detailed and extensive metaphoric knowledge base Martin's approach is highly domain-specific and cannot be applied to other domains without incorporating new handwritten metaphor knowledge. Since we pursue domain-independent metaphor identification, we consider Martin's knowledge-based approach too inflexible for our purposes.

Contrary to Martin's rule-based approach Birke and Sarkar (2006) choose a statistical approach and adapt a word-sense disambiguation method to classify literal and metaphorical occurrences of verbs automatically by clustering techniques. For this purpose they consider the problem of identification of metaphorical language a word-sense disambiguation task between the literal and metaphorical senses of a word. This approach is embodied in their system called TroFi (Trope Finder). However, they emphasize that their system does not interpret metaphors. Instead it only separates literal usages of verbs from non-literal usages. In an evaluation on 25 verbs their system achieves an F-score of 64.9%. Moreover, the authors use the system to compile an example data base (Trope Finder Example Base[1]) for 50 verbs. That data base contains 3737 example sentences where the verb in each sentence is labelled literal or non-literal.

In contrast to Martin's knowledge-based approach Birke & Sarkar show a domain-independent method for metaphor identification. We do not adopt the word-sense-disambiguation aspect of their work because we use frequencies of translations of literal and metaphorical phrases instead. But similar to their system we apply clustering methods to separate literal from non-literal phrases.

Turney et al. (2011) modify the word sense disambiguation approach of Birke and Sarkar (2006) by extending it by Lakoff and Johnson's (1980) hypothesis that views metaphors as a knowledge transfer from a concrete to an abstract domain. Turney et al. (2011) assume the literal or metaphorical sense of a given word to be related to the degree of abstractness of its context: A word is used literally in a concrete context and metaphorically in an abstract context. To calculate the degree of abstractness in a given context an algorithm is used which assigns words

[1] Available at http://www.cs.sfu.ca/~anoop/students/jbirke/.

values between 0 and 1. A value of 0 means that a word is highly concrete while a higher value indicates a higher abstractness. The algorithm is based on the Latent Semantic Analysis (LSA) and generates 114 502 words annotated with abstractness ratings. These abstractness ratings are used to generate feature vectors from a word's context to train a logistic regression model. That model is then applied to new words to assign them the literal or the metaphorical class by means of their context. It it worth noting that for the generation of the word set annotated with abstractness ratings the authors use an initial seed set of 40 concrete and abstract paradigm words. Unlike the authors we do not use any seed set for our approach.

The authors carry out three experiments to evaluate their algorithm. In the first experiment one hundred adjective-noun phrases like deep snow and deep appreciation are labeled literal or metaphorical by five judges. The average classification accuracy achieved here is 79 %. The second and third experiments are performed by means of verbs from the TroFi (Trope Finder) Example Base in different settings. The setting of the second experiment resembles Birke and Sarkar's (2006) setup and achieves an average F-score of 63.9 %. In the third experiment a model is trained on 25 of the 50 TroFi verbs and then tested on the 25 other, previously unseen verbs which is not possible by Birke & Sarkar's approach. Here an F-score of 68.1% is reached.

Turney et al. (2011) achieve remarkable results by combining a word-sense-disambiguation method with an algorithm for abstractness rating. In contrast to Birke and Sarkar's (2006) approach their method is able to classify previously unseen words as well. For our approach we adopt the view that metaphors arise from knowledge transfer from concrete to abstract domains. But contrary to their approach we do not compute abstractness ratings for a target word's context. Instead we investigate to what extent frequencies of translations of a target word are an indication of a metaphorical usage of that target word.

6 Conclusion and Future Work

In this thesis we introduced a novel method for separating literal from metaphorical adjective-noun phrases in English. In contrast to other related approaches our method is completely unsupervised and not restricted to any special domain. Furthermore, it does not rely on any kind of seed set. For this purpose we developed a set of statistical features in order to be able to formally describe adjective-noun phrases. These features include frequency counts and association measures. They were then used to separate 100 English literal and metaphorical adjective-noun phrases into 2 clusters using the K-means clustering algorithm. Furthermore, we delivered a detailed description of our implementation of this method. An extensive evaluation of the method was performed as well.

Our results clearly show that features extracted from translations of adjective-noun phrases can help to identify English metaphors in text. But in contrast to our initial hypothesis which assumed that translations of metaphors would mostly fail we found out that this is not always the case. Instead, the association features computed for the original adjective-noun phrases and their translations proved to be better features than the frequencies of the original adjective-noun phrases and their translations. Moreover, the association features benefit in particular from parallel corpora. The fact that calculating averages of feature values of translation candidates results in scores which are significantly above the baseline is a surprising finding.

The work presented here could be improved and extended in various ways. The resources we used for the experiments are first to mention here. Parallel corpora generally seemed to yield better results but we cannot completely exclude the possibility that the choice of languages also played a role. Therefore a repetition of the experiments using comparable corpora with equal size might bring clarity about the reason for the observed differences. Furthermore, the addition of corpora for other languages and language families like Swedish, Turkish etc. could also lead to new findings. Another shortcoming could be hidden in the size and structure of our metaphor test set. An extension of the test set with further adjective-noun pairs and a greater variety of adjectives could lead to an improvement of the clustering results by means of a larger data base. Finally, the heuristic for the selection of the best translation of an adjective-noun phrase is very simple and therefore still not fully optimized. A deeper analysis of the translation alternatives such as a more elaborate comparison of each translation's features with the features of the

source language phrase could retrieve more appropriate translation candidates. This way a possible pre-processing module could compare the frequencies of the translation candidates' adjectives and nouns to the frequencies of the adjective and the noun of the source language phrase. Using this information a network of adjectives and nouns could be constructed which could then help to improve the selection of translation candidates.

Another directions for future work relate to the used features and the clustering method. The set of features used for this work could be extended by other features, such as the Mutual Information (MI). In this work we use the Pointwise Mutual Information (PMI) which calculates the association between two single elements. Mutual Information, in turn, calculates the sum of association scores between two sets of elements. Therefore the Mutual Information could be an interesting feature to capture the association of all translation candidates for a particular source language phrase. Besides the association measures it is also appropriate to take the contexts of the adjectives and the nouns more into account. In this work we merely considered the numbers of different nouns adjectives can occur with and vice versa. A potential future work could also take the co-occurrent nouns/adjectives directly into account by viewing which co-occurrent adjectives/-nouns are shared by different phrases. This could result in other interesting features.

A clustering into 3 or more different clusters could yield more concise clusters. It is even possible that a greater number of clusters would reveal special kinds of metaphors, for example the conventional metaphors. It is also conceivable that particular metaphorical analogies across languages might reveal more conceptual metaphors like those mentioned by Lakoff and Johnson (1980). This could provide additional hints for the ways in which human perception and cognition work.

Finally, the existing system for metaphor identification might be extended to a system for metaphor interpretation. A metaphor interpretation system paraphrases metaphorical expressions and substitutes figuratively used words by more appropriate words. In the following example from Shutova et al. (2012) the figuratively used verb *swallow* is substituted by the more appropriate *suppress*: *to swallow anger* $\Rightarrow$ *to suppress anger*. The use of translations could help to find such paraphrases. Applications in Computational Linguistics like Machine Translation or Sentiment Analysis could in particular benefit from better metaphor identification and interpretation.

Bibliography

Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, pages 329–336, 2006.

William B. Dolan. Metaphor as an emergent property of machine-readable dictionaries. In *Proceedings of the AAAI 1995 Spring Symposium Series: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, pages 27–32, 1995.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, USA, 1st edition, 2010.

George Lakoff and Mark Johnson. *Metaphors we Live by*. University of Chicago Press, Chicago, 1980.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

James H. Martin. Computer understanding of conventional metaphoric language. *Cognitive Science*, 16(2):233–270, 1992.

Tony McEnery. Corpus linguistics. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, pages 448–463. Oxford University Press, 2003.

Gerard Salton, Andrew Wong, and Chung Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613 – 620, 1975.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing, 1994.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353, 2012.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, 2011.