Linguistic Cues for Distinguishing Literal and Non-Literal Usages

Linlin Li and **Caroline Sporleder**
Department of Computational Linguistics
Saarland University
{linlin, csporled}@coli.uni-saarland.de

Abstract

We investigate the effectiveness of different linguistic cues for distinguishing literal and non-literal usages of potentially idiomatic expressions. We focus specifically on features that generalize across different target expressions. While idioms on the whole are frequent, instances of each particular expression can be relatively infrequent and it will often not be feasible to extract and annotate a sufficient number of examples for each expression one might want to disambiguate. We experimented with a number of different features and found that features encoding lexical cohesion as well as some syntactic features can generalize well across idioms.

1 Introduction

Nonliteral expressions are a major challenge in NLP because they are (i) fairly frequent and (ii) often behave idiosyncratically. Apart from typically being semantically more or less opaque, they can also disobey grammatical constraints (e.g., *by and large*, *lie in wait*). Hence, idiomatic expressions are not only a problem for semantic analysis but can also have a negative effect on other NLP applications (Sag et al., 2001), such as parsing (Baldwin et al., 2004).

To process non-literal language correctly, NLP systems need to recognise such expressions automatically. While there has been a significant body of work on idiom (and more generally multiword expression) detection (see Section 2), until recently most approaches have focused on a *type-based classification*, dividing expressions into "idiomatic" or "not idiomatic" irrespective of their actual use in a discourse context. However, while some expressions, such as *by and large*, always have a non-compositional, idiomatic meaning, many other expressions, such as *break the ice* or *spill the beans*, can be used literally as well as idiomatically and for some expressions, such as *drop the ball*, the literal usage can even dominate in some domains. Consequently, those expressions have to be disambiguated in context (*token-based classification*).

We investigate how well models for distinguishing literal and non-literal use can be learned from annotated examples. We explore different types of features, such as the local and global context, syntactic properties of the local context, the form of the expression itself and properties relating to the cohesive structure of the discourse. We show that several feature types work well for this task. However, some features can generalize across specific idioms, for instance features which compute how well an idiom "fits" its surrounding context under a literal or non-literal interpretation. This property is an advantage because such features are not restricted to training data for a specific target expression but can also benefit from data for other idioms. This is important because, while idioms as a general linguistic class are relatively frequent, instances of each particular idiom are much more difficult to find in sufficient numbers. The situation is exacerbated by the fact the distributions of literal vs. non-literal usage tend to be highly skewed, with one usage (often the non-literal one) being much more frequent than the other. Finding sufficient examples of the minority class can then be difficult, even if instances are extracted from large corpora. Furthermore, for highly skewed distributions, many more majority class examples have to be annotated to obtain an acceptable number of minority class instances.

We show that it is possible to circumvent this problem by employing a generic feature space that

Coling 2010: Poster Volume, pages 683–691,
Beijing, August 2010

looks at the cohesive ties between the potential idiom and its surrounding discourse. Such features generalize well across different expressions and lead to acceptable performance even on expressions unseen in the training set.

2 Related Work

Until recently, most studies on idiom classification focus on type-based classification; sofar there are only comparably few studies on token-based classification. Among the earliest studies on token-based classification were the ones by Hashimoto et al. (2006) on Japanese and Katz and Giesbrecht (2006) on German. Hashimoto et al. (2006) present a rule-based system in which lexico-syntactic features of different idioms are hard-coded in a lexicon and then used to distinguish literal and non-literal usages. The features encode information about the passivisation, argument movement, and the ability of the target expression to be negated or modified. Katz and Giesbrecht (2006) compute meaning vectors for literal and non-literal examples in the training set and then classify test instances based on the closeness of their meaning vectors to those of the training examples. This approach was later extended by Diab and Krishna (2009), who take a larger context into account when computing the feature vectors (e.g., the whole paragraph) and who also include prepositions and determiners in addition to content words.

Cook et al. (2007) and Fazly et al. (2009) take a different approach, which crucially relies on the concept of *canonical form* (CForm). It is assumed that for each idiom there is a fixed form (or a small set of those) corresponding to the syntactic pattern(s) in which the idiom normally occurs (Riehemann, 2001).The canonical form allows for inflectional variation of the head verb but not for other variations (such as nominal inflection, choice of determiner etc.). It has been observed that if an expression is used idiomatically, it typically occurs in its canonical form (Riehemann, 2001). Cook et al. exploit this behaviour and propose an unsupervised method in which an expression is classified as idiomatic if it occurs in canonical form and literal otherwise. Canonical forms are determined automatically using a statistical, frequency-based measure.

Birke and Sarkar (2006) model literal vs. non-literal classification as a word sense disambiguation task and use a clustering algorithm which compares test instances to two seed sets (one with literal and one with non-literal expressions), assigning the label of the closest set.

Sporleder and Li (2009) propose another unsupervised method which detects the presence or absence of cohesive links between the component words of the idiom and the surrounding discourse. If such links can be found the expression is classified as literal otherwise as non-literal. Li and Sporleder (2009) later extended this work by combining the unsupervised classifier with a second-stage supervised classifier.

Hashimoto and Kawahara (2008) present a supervised approach to token-based idiom distinction for Japanese, in which they implement several features, such as features known from other word sense disambiguation tasks (e.g., collocations) and idiom-specific features taken from Hashimoto et al. (2006). Finally, Boukobza and Rappoport (2009) also experimented with a supervised classifier, which takes into account various surface features.

In the present work, we also investigate supervised models for token-based idiom detection. We are specifically interested in which types of features (e.g., local context, global context, syntactic properties) perform best on this task and more specifically which features generalize across idioms.

3 Data

We used the data set created by Sporleder and Li (2009), which consists of 13 English expressions (mainly V+PP or V+NP) that can be used both literally and idiomatically, such as *break the ice* or *play with fire*.[1] To create the data set all instances of the target expressions were extracted from the Gigaword corpus together with five paragraphs of context and then labelled manually as 'literal' or 'non-literal'. Overall the data set consists of just under 4,000 instances. For most ex-

[1] We excluded four expressions from the original data set because their number of literal examples was very small (< 2).

pressions the distribution is heavily skewed towards the idiomatic interpretation, however for some, like *drop the ball*, the literal reading is more frequent. The number of instances varies, ranging from 15 for *pull the trigger* to 903 for *drop the ball.* While the instances were extracted from a news corpus, none of them are domain-specific and all expressions also occur in the BNC, which is a balanced, multi-domain corpus.

To compute the features which we extract in the next section, all instances in our data sets were part-of-speech tagged by the MXPOST tagger (Ratnaparkhi, 1996), parsed with the MaltParser[2], and named entity tagged with the Stanford NE tagger (Finkel et al., 2005). The lemmatization was done by RASP (Briscoe and Carroll, 2006).

4 Indicators of Idiomatic and Literal Usage

In this study we are particularly interested in which linguistic indicators work well for the task of distinguishing literal and idiomatic language use. The few previous studies have mainly looked at the lexical context in which and expression occurs (Katz and Giesbrecht, 2006; Birke and Sarkar, 2006). However, other properties of the linguistic context might also be useful. We distinguish these features into different groups and discuss them in the following sections.

4.1 Global Lexical Context (glc)

That the lexical context might be a good indicator for the usage of an expression is obvious when one looks at examples as in (1) and (2), which suggest that literal and non-literal usages of a specific idiom co-occur with different sets of words. Non-literal uses of *break the ice* (1), for instance, tend to occur with words like *discuss*, *bilateral* or *relations*, while literal usages (2) predictably occur with, among others, *frozen*, *cold* or *water*. What we are looking at here is the global lexical context of an expression, i.e., taking into account previous and following sentences. We are specifically looking for words which are either semantically related (in a wide sense) to the literal or the non-literal sense of the target expression. The presence or absence of such words can be a good indicator of how the expression is used in a context.

(1) ”Gujral will meet Sharif on Monday and **discuss bilateral relations**,” the Press Trust of India added. The minister said Sharif and Gujral would be able to ”break the ice” over Kashmir.

(2) Meanwhile in Germany, the **cold** penetrated Cologne cathedral, where worshippers had to break the ice on the **frozen** holy **water** in the font.

We implemented two sets of features which encode the global lexical context: *salient words* and *related words* as described in Li and Sporleder (2009). The former feature uses a variant of tf.idf to identify words that are particulary salient for different usages. The latter feature identifies words which are most strongly related to the component words of the idiom.

We notice that sometimes several idioms co-occur within the same instance. This is to say that nonliteral usages may be indicators of each other since authors may put them in a same context to convey a specific opinion (e.g., irony). Due to this, global lexical context features may also generalize across idioms to some extend.

4.2 Local Lexical Context (locCont)

In addition to the global context, the local lexical context, i.e., the words preceding and following the target expression, might also provide important information. One obvious local clue are words like *literally* or *metaphorically speaking*, which when preceding or following an expression might indicate its usage. Unfortunately, such clues are not only very rare (we only found a handful in nearly 4,000 annotated examples) but also not always reliable. For instance, it is not difficult to find examples like (3) and (4) where the word *literally* is used even though the idiom clearly has a non-literal meaning.

(3) In the documentary the producer **literally** spills the beans on the real deal behind the movie production.

(4) The new philosophy is blatantly permissive and **literally** passes the buck to the House’s other committees.

[2] `http://maltparser.org/index.html`

However, there are other local cues. For example, we found that the word *just* before *get ones feet wet* tends to indicate non-literal usage as in (5). Non-literal usage can also be indicated by the occurrence of the prepositions *over* or *between* after *break the ice* as in (1) and (6). While such cues are not perfect they often make one usage more likely than the other. Unlike the semantically based global cues, many local clues are more rooted in syntax, i.e., local cues work because specific *constructions* tend to be more frequent for one or the other usage.

(5) The wiki includes a page of tasks suitable for those **just** getting their feet wet.

(6) Would the visit of the minister help break the ice **between** India and Pakistan?

Another type of local cues involves selectional preferences. For example, idiomatic usage is probable if the subject of *play with fire* is a country as in (7) or if *break the ice* is followed by a *with*-PP whose NP refers to a person (8).

(7) Dudayev repeated his frequent warnings that **Russia** was playing with fire.

(8) Edwards usually manages to break the ice with the taciturn **monarch**.

Based on those observations, we encode which words occur in a ten word window around the target expression, five pre-target words and five post-target words, as the locCont features.

4.3 Discourse Cohesion (dc)

We implemented two features, *related score* and *discourse connectivity*, which take into account the cohesive structure of an expression in its context as described by Li and Sporleder (2009). In addition, we also included the prediction of the cohesion graph proposed by Sporleder and Li (2009) as an additional feature. These features look at the lexical cohesion between an expression and the surrounding discourse, so they are more likely to generalize across different idioms.

4.4 Syntactic Structure (allSyn)

To capture syntactic effects, we encoded information of the **head node (heaSyn)** of the target expression in the dependency tree (e.g., *break* in the dependency tree in Figure 1). The syntactic features we encoded are the **parent node (parSyn)**, **sibling nodes (sibSyn)** and **children nodes (chiSyn)** of the head node. These nodes include the following type of syntactic information:

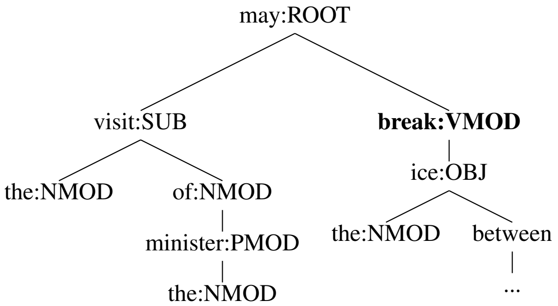

Figure 1: Dependency tree for a nonliteral example of *break the ice* (*The visit of the minister may break the ice between India and Pakistan.*)

Dependency Relation of the Verb Phrase The whole idiomatic expression used as an object of a preposition can be an indicative factor of idiomatic usage (see Example 9). This property is captured by the *heaSyn* feature.

(9) Ross headed back last week to Washington to brief president Bill Clinton on the Hebron talks after achieving a breakthrough **in** breaking the ice in the Hebron talks by arranging an Arafat-Netanyahu summit .

Modal Verbs usually appear in the parent position of the head verb (*parSyn*). Modals can be an indicator of idiomatic usage such as **may** in Figure 1. In contrast, the modal **had to** is indicative that the same phrase is used literally (Example 10).

(10) Dad **had to** break the ice on the chicken troughs.

Subjects can also provide clues about the usage of an expression, e.g., if selectional preferences are disobeyed. For instance, **visit** as a subject of the verb phrase *break the ice* is an indicator of idiomatic usage (see Figure 1). Subjects typically appear in the children position of the head verb (*chiSyn*), but sometimes may appear in the sibling position (*sibSyn*) as in Figure 1 .

Verb Subcat We also encode the arguments of the head verb of the target expression. These arguments can be, for example, additional PPs. This feature encodes syntactic constraints and attempts to model selectional restrictions. The likelihood of subcategorisation frames may differ for the two usages of an expression, e.g., non-literal expressions often tend to have a shorter argument list. For instance, the subcat frame <PP-on, PP-for> intuitively seems more likely for literal usages of the expression *drop the ball* (see Example 11) than for non-literal ones, for which <PP-on> is more likely (12). To capture subcategorisation behaviour, we encode the children nodes of the head node (*chiSyn*).

(11) US defender Alexi Lalas twice went close to forcing an equaliser , first with a glancing equaliser from a Paul Caligiuri free kick and then from a Wynalda corner when Prunea dropped the ball **[on the ground]** only **[for Tibor Selyme to kick frantically clear]** .

(12) "Clinton dropped the ball **[on this]**," said John Parachini.

Modifiers of the verb can also be indicative of the usage of the target expression. For example, in 13, the fact that the phrase *get one's feet wet* is modified by the adverb *just* suggest that it is used idiomatically. Similar to verb subcat, modifiers are often appear in the children position (*chiSyn*).

(13) The wiki includes a page of tasks suitable for those **just** getting their feet wet.

Coordinated Verb Which verbs are coordinated with the target expression, if any, can also provide cues for the intended interpretation. For example, in (14), the fact that *break the ice* is coordinated with *fall* suggest that it is used literally. The coordinated verb can appear at the sibling position, children position, or some other position of the head verb depending on the parser. The Malt-parser tends to put the coordinated verbs in the children position (*chiSyn*).

(14) They may break the ice and **fall** through.

4.5 Other Features

Named Entities (ne) can also indicate the usage of an expression. For instance, a country name in the subject position of the target expression *break the ice* is a strong indicator of this phrase being used idiomatically (see Example 7). Diab and Bhutada (2009) find that NE-features perform best. They used a commercial NE-tagger with 19 classes. We used the Stanford NE tagger (Finkel et al., 2005), and encoded three named entity classes ("person", "location", "organiszation") in the feature vector.

Indicative Terms (iTerm) Some words such as *literally, proverbially* are also indicative of literal or idiomatic usages. We encoded the frequencies of those indicative terms as features.

Scare Quotes (quote) This feature encodes whether the idiom is marked off by scare quotes, which often indicates non-literal usage (15).

(15) Do consider "getting your feet wet" online, using some of the technology that is now available to us.

5 Experiments

In the previous section we discussed different linguistic cues for idiom usage. To determine which of these cues work best for the task and which ones generalize across different idioms, we carried out three experiments. In the first one (Section 5.1) we trained one model for each idiom (see Section 3) and tested the predictiveness of each feature type individually as well as all features together. In the second experiment (Section 5.2), we trained one generic model for all idioms and determined how the performance of this model differs from the idiom-specific models. Specifically we wanted to know whether the model would benefit from the additional training data available by combining information from several idioms. Finally (Section 5.3), we tested the generic model on *unseen* idioms to determine whether these could be classified based on generic properties even if training data for the target expressions had not been seen.

5.1 Idiom Specific Models

The first question we wanted to answer was how difficult token-based idiom classification is and which of the features we defined in the previous section work well for this task. We implemented a specific classifier for each of the idioms in the data set. We trained one model for all features in combination and one for each individual feature. Because the data set is not very big we decided to run these experiments in 10-fold stratified cross-validation mode. We used the SVM classifier (SMO) from Weka.[3]

Table 1 shows the results. We report the precision (Prec.), recall (Rec.) and F-Score for the literal class, as well as the accuracy. Note that due to the imbalance in the data set, accuracy is not a very informative measure here; a classifier always predicting the majority class would already obtain a relatively high accuracy. The literal F-Score obtained for individual idioms varies from 38.10% for *bite one's tongue* to 96.10% for *bounce of the wall*. However, the data sets for the different idioms are relatively small and it is impossible to say whether performance differences on individual idioms are accidental, or due to differences in training set size or due to some inherent difficulty of the individual idiom. Thus we chose not to report the performance of our models on individual idioms but on the whole data set for which the numbers are much more reliable. The final performance confusion matrix is the sum over all individual idiom confusion matrices.

	Avg. literal			Avg.
feature	Prec.	Rec.	F-Score	Acc.
all	89.84	77.06	82.96	93.36
glc+dc	90.42	76.44	82.85	93.36
allSyn	76.30	86.13	80.92	91.48
heaSyn	76.64	85.77	80.95	91.53
parSyn	76.43	88.34	81.96	91.84
chiSyn	76.49	88.22	81.94	91.84
sibSyn	76.27	88.34	81.86	91.78
locCont	76.51	88.34	82.00	91.86
ne	76.49	88.22	81.94	91.84
iTerm	76.51	88.34	82.00	91.86
quote	76.51	88.34	82.00	91.86
Base_{maj}	76.71	88.34	82.00	91.86

Table 1: Performance of idiom-specific models (averaged over different idioms), 10-fold stratified cross-validation.

The Baseline (Base) is built based on predicting the majority class for each expression. This means predicting *literal* for the expressions which consist of more literal examples and *nonliteral* for the expressions consisting of more nonliteral examples. We notice the baseline gets a fairly high performance (Acc.=91.86%).

The results show that the expressions can be classified relatively reliably by the proposed features. The performance beats the majority baseline statistically significantly ($p = 0.01$, χ^2 test). We noticed that parSyn, chiSyn, locCont, iTerm and quote features are too sparse. These individual features cannot guide the classifier. As a result, the classifier only predicts the majority class which results in a performance similar to the baseline. Some of the syntactic features are less sparse and they get different results from the baseline classifier, however, the performances of these features are actually worse than the baseline. This may be due to the relatively small training size in each idiom specific model. When adding those features together with statistical-based features (glc+dc), the performance of the literal class can be improved slightly. However, we did not observe any performance increase on the accuracy.

5.2 Generic Models

Having verified that literal and idiomatic usages can be distinguished with some success by training expression-specific models, we carried out a second experiment in which we merged the data sets for different expressions and trained one generic model. We wanted to see whether a generic model, which has access to more training data, performs better and whether some features, e.g., the cohesion features profit more from this. The experiment was again run in 10-fold stratified cross-validation mode (using 10% from each idiom in the test set in each fold).

Table 2 shows the results. The baseline classifier always predict the majority class 'nonliteral'. Note that the result of this baseline is different from the majority baseline in the idiom specific model. In the idiom specific model, there are three expressions [4] for which the majority class is 'literal'.

Unsurprisingly, the F-Score and accuracy of the combined feature set drops a bit. However, the performance still statistically significantly beats the majority baseline classifier ($p << 0.01$, χ^2 test). Similar to previous observation, the

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

[4] I.e., *bounce off the wall*, *drop the ball*, *pull the trigger*

	Avg. literal			Avg.
feature	Prec.	Rec.	F-Score	Acc.
all	89.59	65.77	73.22	89.90
glc+dc	82.53	60.86	70.06	89.08
allSyn	50.83	59.88	54.99	79.42
heaSyn	50.57	59.88	54.83	79.29
sibSyn	33.33	0.86	1.67	78.83
ne	62.45	20.00	30.30	80.69
iTerm	40.00	0.25	0.49	78.99
Base$_{maj}$	–	–	–	79.01

Table 2: Performance of the generic model (averaged over different idioms), 10-fold stratified cross-validation.

statistical-based features (glc+dc) work the best, while the syntactic features are also helpful. However, the local context, iTerm, quote features are very sparse and, as in the idiom-specific experiments, the performances of these features are similar to the majority baseline classifier. We excluded them from the Table 2.

The numbers show that the syntactic features help more in this model compared with the idiom-specific model. When including these features, literal F-Score increases by 3.16% while accuracy increases by 0.9%. It seems that the syntactic features benefit from the increased training set. This is evidence that these features can generalize across idioms. For instance, the phrase "The US" on the subject position may be not only indicative of the idiomatic usage of *break the ice*, but also of idiomatic usage of *drop the ball.*

We found that the indicative terms are rare in our corpus. This is the reason why the recall rate of the indicative terms is very low (0.25%). The indicative terms are not very predictive of literal or non-literal usage, since the precision rate is also relatively low (40%), which means those words can be used in both literal and nonliteral cases.

5.3 Unseen Idioms

In our final experiment, we tested whether a generic model can also be applied to completely new expressions, i.e., expressions for which no instances have been seen in the data set. Such a behaviour would be desireable for practical purposes as it is unrealistic to label training data for each idiom the model might possibly encounter in a text. To test whether the generic model does indeed generalize to unseen expressions, we test it on all instances of a given expression while training on the rest of the expressions in the dataset. That is, we used a modified cross-validation setting, in which each fold contains instances from one expression in the test set. Since our dataset contains 13 expressions, we run a 13-fold cross validation. The final confusion matrix is the sum over each confusion matrix in each round.

	Avg. literal			Avg.
feature	Prec.	Rec.	F-Score	Acc.
all	96.70	81.65	88.54	95.41
glc+dc	96.93	77.00	85.83	94.48
allSyn	52.54	58.77	55.48	79.52
heaSyn	51.35	59.47	55.11	78.96
sibSyn	55.56	2.32	4.46	78.38
ne	61.89	19.05	29.13	79.87
iTerm	66.67	0.7	1.38	78.36
Base$_{maj}$	–	–	–	79.01

Table 3: Performance of the generic model on unseen idioms (cross validation, instances from each idiom are chosen as test set for each fold)

The results are shown in Table 3. Similar to the generic model, we found that the cohesion features and syntactic features do generalize across expressions. Statistical features (glc+dc) perform well in this experiment. When including more linguistically orientated features, the performance can be further increased by almost 1%. In line with former observations, the sparse features mentioned in the former two experiments also do not work for this experiments. We also excluded them from the table.

One interesting finding about this experiment of this model is that the F-Score is higher than for the "generic model". This is counter-intuitive, since in the generic model, each idiom in the testing set has examples in the training set, thus, we might expect the performance to be better due to the fact that instances from the same expression appearing in the training set are more informative compared with instances from different idioms. Further analysis revealed that there are some expressions for which it may actually be beneficial to train on other expressions, as the evidence of some features may be misleading.

	literal F-S.		Acc.	
feature	Spe.	Gen.	Spe.	Gen.
all	86.85	**91.79**	80.67	**88.37**
glc+dc	86.75	**88.84**	80.67	**84.61**
allSyn	**85.71**	71.94	**75.28**	61.13
heaSyn	**85.79**	71.94	**75.39**	61.13

Table 4: Comparing the performance of the idiom *drop the ball* on the idiom specific model (Spe.) and generic model (Gen.)

Table 4 shows the comparison of the performance of *drop the ball* on the idiom specific model and the generic model on unseen idioms. It can be seen that the statistical features (glc+dc) work better for the model that is trained on the instances from other idioms than the model which is trained on the instances of the target expression itself. We found this is due to the fact that *drop the ball* is especially difficult to classify with the discourse cohesion features (dc). The literal cases are often found in a context containing words, such as **fault**, **mistake**, **fail**, and **miss**, which are used to describe a scenario in a baseball game,[5] while, on the other hand, those context words are also closely semantically related to the idiomatic reading of *drop the ball*. This means the classifier can be mislead by the cohesion features of the literal instances of this idiom in the training set, since they exhibit strong idiomatic cohesive links with the target expression. When excluding *drop the ball* from the training set, the cohesive links in the training data are less noisy. Thus, the performance increases. Unsurprisingly, the performance of syntactic features works better for the idiom specific model compared with the unseen idiom model.

6 Conclusion

Idioms on the whole are frequent but instances of each particular idiom can be relatively infrequent (even for common idioms like “spill the beans”). The classes can also be fairly imbalanced, with one class (typically the nonliteral interpretation) being much more frequent than the other. This causes problems for training data generation. For idiom specific classifiers, it is difficult to obtain large data sets even when extracting from large corpora and it is even more difficult to find sufficient examples of the minority class. In order to address this problem, we looked for features which can generalize across idioms.

We found that statistical features (glc+dc) work best for distinguishing literal and nonliteral readings. Certain linguistically motivated features can further boost the performance. However, those linguistic features are more likely to suffer from data sparseness, as a result, they often only predict the majority class if used on their own. We also found that some of the features that we designed generalize well across idioms. The cohesion features have the best generalization ability, while syntactic features can also generalize to some extent.

Acknowledgments

This work was funded by the DFG within the Cluster of Excellence MMCI.

References

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephen Oepen. 2004. Road-testing the English resource grammar over the British National Corpus. In *Proc. LREC-04*, pages 2047–2050.

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

Boukobza, Ram and Ari Rappoport. 2009. Multiword expression identification using sentence surface features. In *Proceedings of EMNLP-09*.

Briscoe, Ted and John Carroll. 2006. Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 41–48.

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

[5] The corpus contains many sports news text

Diab, Mona and Pravin Bhutada. 2009. Verb noun construction mwe token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22.

Diab, Mona T. and Madhav Krishna. 2009. Unsupervised classification of verb noun multi-word expression tokens. In *CICLing 2009*, pages 98–110.

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

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL-05*, pages 363–370.

Hashimoto, Chikara and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of EMNLP-08*, pages 992–1001.

Hashimoto, Chikara, Satoshi Sato, and Takehito Utsuro. 2006. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of COLING/ACL-06*, pages 353–360.

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

Li, Linlin and Caroline Sporleder. 2009. Contextual idiom detection without labelled data. In *Proceedings of EMNLP-09*.

Ratnaparkhi, Adwait. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of EMNLP-96*.

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

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: a pain in the neck for NLP. In *Lecture Notes in Computer Science*.

Sporleder, Caroline and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL-09*.