Idioms in Context: The IDIX Corpus

Caroline Sporleder, Linlin Li, Philip John Gorinski, Xaver Koch

Computational Linguistics / Cluster of Excellence MMCI
Saarland University
{csporled,linlin,philipg,xkoch}@coli.uni-saarland.de

Abstract

Idioms and other figuratively used expressions pose considerable problems to natural language processing applications because they are very frequent and often behave idiosyncratically. Consequently, there has been much research on the automatic detection and extraction of idiomatic expressions. Most studies focus on type-based idiom detection, i.e., distinguishing whether a given expression can (potentially) be used idiomatically. However, many expressions such as *break the ice* can have both literal and non-literal readings and need to be disambiguated in a given context (token-based detection). So far relatively few approaches have attempted context-based idiom detection. One reason for this may be that few annotated resources are available that disambiguate expressions in context. With the IDIX corpus, we aim to address this. IDIX is available as an add-on to the BNC and disambiguates different usages of a subset of idioms. We believe that this resource will be useful both for linguistic and computational linguistic studies.

1. Introduction

Idioms are multi-word expressions whose meaning cannot be inferred from the meaning of their parts in a completely compositional manner. As a class, idioms and other figurative expressions are relatively frequent. For example, Burchardt et al. (2006) found that nearly 15% of all verb occurrences in a German newspaper corpus are used figuratively, rising to nearly 83% for high frequency verbs like *nehmen* (to take).

Idioms tend to behave idiosyncratically, not only with respect to their semantics but also with respect to other linguistic properties. For example, they typically exhibit some degree of lexical and syntactic fixedness (e.g., *raining cats and dogs* vs. **raining cats and hounds* or **raining dogs and cats*). They can also be syntactically anomalous (e.g., *in line* without a determiner in front of *in*), violate selectional restrictions (as in *push one's luck* under the assumption that only concrete things can normally be pushed), or change the default assignment of semantic roles to syntactic categories (e.g., in *break sth with X*, the argument X would typically be an instrument but for the idiom *break the ice* it is more likely to fill a patient role, as in *break the ice with Syria*).

Such anomalies can cause significant problems to natural language processing (NLP) tools, and the situation is considerably worsened by the fact that idioms occur frequently in natural texts. For instance, Baldwin et al. (2004) found that 8% of all parse failures obtained with the English Resource Grammar could be attributed to idioms and other multi-word expressions. Proper recognition and modelling of idioms has also been shown to benefit other applications (Lin, 1998; Gerber and Yang, 1997; Bond and Shirai, 1997; Lewis and Croft, 1990).

Hence, being able to recognise idioms is crucial for NLP applications. However, this is not a trivial task. Machine readable idiom dictionaries could help to some extent but usually lack coverage. To alleviate this problem, several studies have addressed the question of whether idiom lists can be compiled automatically from text corpora, e.g., by employing statistical measures to assess the idiomaticity of a phrase (so-called *type-based idiom classification* (Bannard, 2007; Fazly and Stevenson, 2006; Baldwin et al., 2003; Lin, 1999)).

While there has been some success in this direction, type-based idiom classification only solves part of the problem; many expressions can have a literal as well as an idiomatic meaning (see Examples (1a) vs. (1b) and (2a) vs. (2b))[1] and thus need to be disambiguated in context (*token-based idiom classification*). So far, relatively few studies have addressed this task, though there has been an increased interest recently (Li and Sporleder, 2010; Fazly et al., 2009; Diab and Krishna, 2009a; Diab and Krishna, 2009b; Li and Sporleder, 2009a; Sporleder and Li, 2009; Li and Sporleder, 2009b; Cook et al., 2007; Birke and Sarkar, 2006; Katz and Giesbrecht, 2006; Hashimoto et al., 2006a; Hashimoto et al., 2006b).

One reason why token-based idiom detection has re-

[1]The examples in this paper are taken either from the BNC http://www.natcorp.ox.ac.uk/ or the Gigaword corpus http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05.

ceived less attention than type-based detection in the past is probably the relative lack of freely available data sets in which potentially ambiguous expressions are disambiguated in context. The few available resources include the VNC-Tokens Dataset (Cook et al., 2008), which provides annotations for 53 verb-noun combinations (around 3000 sentences in total) extracted from the BNC. A similar resource for German is discussed in Fritzinger et al. (2010), which contains 77 different preposition-noun-verb triples with around 9,700 annotated instances in total, extracted from EUROPARL (Koehn, 2005) and a newspaper corpus (*Frankfurter Allgemeine Zeitung*). For Japanese, Hashimoto and Kawahara (2008) compiled a data set for 146 idioms with 102,846 sentences. A related resource which disambiguates literal and non-literal usages for individual verbs rather than multi-word expressions is the TroFi Example Base (Birke, 2005). This data sets contains occurrences of 50 verbs from the Wall Street Journal and is partly manually annotated.[2]

(1) a. Dad had to break the ice on the chicken troughs so that they could get water. [GIGA NYT200008]

b. If you've just moved to a new area a good way to break the ice for you and your child is a parent and toddler group. [BNC AAY]

(2) a. Somehow I always end up spilling the beans all over the floor and looking foolish when the clerk comes to sweep them up. [GIGA NYT199712]

b. With the VSX4 test suite already announced and starting to ship, X/Open Co is preparing to go public on XPG4 itself, and has set early October as the time when it will spill the beans. [BNC CTV]

With the creation of the IDIX corpus (IDioms In conteXt) we aim to provide another resource for token-based idiom detection. IDIX will be available as an add-on to the British National Corpus (BNC).[3] It contains annotations for all BNC occurrences of a subset of potentially idiomatic expressions. These occurrences are manually labelled as 'literal', 'idiomatic', 'mixed reading', 'meta-linguistic', 'embedded in a larger figurative expression' and 'undecidable' (see Section 2. for details). The annotation scheme thus goes beyond a binary 'literal' vs. 'non-literal' distinction. The annotation is still ongoing. Currently the corpus contains annotations for 78 expressions (5,836 instances in total) but we aim to extend it to at least 100 expressions.

2. The Corpus

When compiling the corpus we had to address several design questions, which we discuss in this section.

2.1. Corpus Choice

IDIX is made available as an add-on to the BNC XML Edition. Our decision to use data from the BNC was motivated by a number of factors: First, the BNC is a balanced corpus which contains text from various domains and genres. By covering a multitude of domains, we hope that our data will be relatively domain-independent and useful for a number of application scenarios. Furthermore, since the BNC provides information on which texts come from which domain, it will be possible to use IDIX for (small-scale) corpus linguistic studies, e.g., regarding the distribution of literal vs. idiomatic readings of a given idiom in different domains and genres.

Second, the BNC XML Edition is very well preprocessed and already comes with several annotation layers, such as automatically assigned part-of-speech tags and syntactic parse trees produced by the RASP parser (Briscoe et al., 2006). This made it easier for us to process the data and extract examples for annotation. Moreover, we hope these additional annotation layers will also be useful for other researchers who might want to work with the corpus.

On the downside, we are not able to make available the annotated data directly, since the BNC requires a license. Instead, we will release the annotation labels for each example together with its file number and sentence id.[4] This will make it easy to map the annotations back to the BNC sources. Another drawback is that the BNC is relatively small (1 million words) compared to some other corpora, such as the Gigaword corpus (1.7 billion words), or data extracted from the Web. This means that we can only find relatively few instances for each idiom. However, we believe that the advantages of having a clean, multi-domain corpus far outweigh these drawbacks.

[2] See the MWE data repository for a list of available resources for different types of multiword expressions: http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets

[3] http://www.natcorp.ox.ac.uk/

[4] To deal with the unlikely case that a given expression occurs repeatedly in a sentence, we also keep track of where in the sentence an expression occurs. However, so far we did not have to make use of this information.

2.2. Idiom Selection and Extraction

For the time being, we mainly (but not exclusively) annotated expressions of the form V+NP and V+PP. This is a relatively frequent syntactic pattern for idioms and many idioms of this type share their form with a literal usage. When selecting expressions for annotation we proceeded as follows: We first looked for expressions in idiom dictionaries, mainly Cowie et al. (1997). We then used the Google search engine to get a first impression of how frequent the respective idiom is on the internet. Afterwards we looked for contexts containing these expressions via the BNC online search interface and quickly browsed the results to obtain a rough idea of the proportion of literal vs. non-literal readings. We favoured expressions which are (i) relatively frequent in the BNC/on the Internet, and (ii) can be found with both literal and idiomatic meanings in the BNC.

Finally, we automatically extracted all occurrences of the target expressions from the parsed version of the BNC[5] using a Perl script that is able to handle search queries on the command line.

These queries impose restrictions on the parse trees of candidate sentences. For example, (3) shows the query used to extract occurrences of the expression *get cold feet*, where "=" indicates a dependency relation. The query matches sentences in which the noun *foot* is directly dependent on the verb *get*, and the adjective *cold* is directly dependent on *foot*. Two extracted sentences are given in (4) and (5) along with their corresponding dependency trees in Figures 1 and 2.

(3) get:VERB=foot:SUBST;
foot:SUBST=cold:ADJ

(4) He gets cold feet and phones his bank manager asking him to stop the cheque. [BNC C8V]

(5) 'I wonder why people get cold feet,' Killion remarked. [BNC HRA]

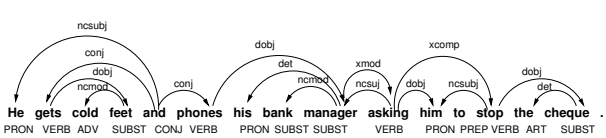

Figure 1: RASP Dependency Tree for Example (4)

Working with dependencies on the parse trees allowed us to cover a wide variety of linguistic forms. In particular, we included also so-called non-canonical

[5]We used the official RASP (Briscoe et al., 2006) parses released with the BNC XML Edition.

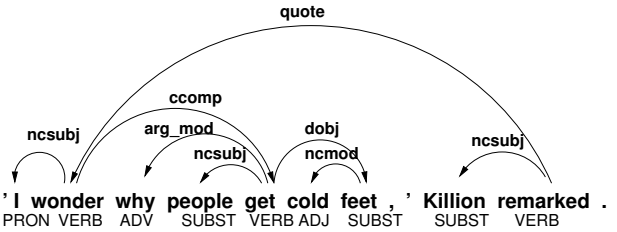

Figure 2: RASP Dependency Tree for Example (5)

forms, i.e., syntactic and lexical variations of the dictionary form that go beyond verb inflection. For instance, we were also able to extract realisations of idioms where the comprising words are split up (6), resulting in a very long distance between them otherwise not easily covered, or expressions with an altered word order (7) (both examples are for the expression *answer the call (of duty))*.

(6) Helmut Kohl's calls to brief Major on the British economy are doubtless answered by a builder screaming at the German Chancellor to speak English, or ordering him to get off the line so they can deal with the other jobs they've got on the go. [BNC CH1]

(7) Another call was answered by Mr Holt on February 20. [BNC E9U]

Of course, this strategy makes the extraction process somewhat dependent on the correctness of the syntactic structures, potentially leading to false negatives (or positives) if a sentence containing an instance of the target idiom has been parsed incorrectly. However, we found that for most expressions it is enough to only specify one dependency (e.g., between the verb and the noun in the target expression) to obtain good retrieval results. Consequently, the retrieval method is relatively robust against parser errors.

Within the dependency-query, wildcards replacing words and POS-tags can be used, providing the possibility to search for idioms where the specific realisation of a word is unclear, e.g., *hold somebody's hand*, where the possessor can be realised by a number of pronouns, including *his*, *her*, or *my*. Searching with a wildcard instead of a word (8) will extract all the instances where the wildcard's part of speech is filled by any word.

(8) hold:VERB=hand:SUBST;
hand:SUBST=?:PRON

The extraction script works in a greedy manner, i.e., each and every sentence that fulfills the given dependencies is extracted as an instance of the target expression. While this approach works well for the majority of the targeted expressions, there are some cases

where it leads to a lot of overhead in the form of erroneous extractions.
For example, query (8) is not general enough to extract all instances of the expression *hold somebody's hand*, as the possessor cannot only be realised by a pronoun but also by a possessive noun (see example (9)). Thus, the correct dependency query has to take into account that the possessor can be tagged as PRON as well as SUBST, making it necessary to use another wildcard (10) or to completely omit the dependency of the patient (11).

(9) He was holding Mary's hand.

(10) hold:VERB=hand:SUBST; hand:SUBST=?:?

(11) hold:VERB=hand:SUBST

We decided on using queries like (11) when necessary in the extraction process, typically extracting all examples with a dependency between the verb and the noun in the expression. Erroneous extractions were then manually filtered out in the annotation process. This high-recall strategy ensures that we can be reasonably sure that we catch most if not all occurrences of the target expressions, i.e., we will get a (near) perfect coverage for the target expression. For later reproducibility, the queries used for the extractions were recorded.

2.3. Annotation Process

For annotation, we implemented a tool, SAI (Simple Annotator for IDIX), that provides a graphic user interface to the annotator, displaying pure text without XML-tags (see Figure 3).
SAI allows the annotator to see the instance to be annotated embedded in context (two paragraphs before and after the occurrence of the instance, with the target expression highlighted). It is possible to switch back and forth between the extracted instances and change the preset default label if necessary. This procedure ensured that the annotator could fully concentrate on the task of labeling the target expressions, without being distracted by meta-information during the annotation process.
To ensure high annotation quality, regular meetings were held in order to discuss difficult cases, e.g., instances with unclear meanings. These cases were also recorded in an annotation Wiki for later reference.

2.4. Annotation Labels and Guidelines

When we started with the annotation process, we intended to use four labels: 'literal' ('l'), 'non-literal' ('n'), 'unclear/undecidable' ('?') and 'false extraction' ('f'). However, it soon became clear that the situation was more complicated. First, we found a few examples in which the target expression was used meta-linguistically ('m') as in (12). Second and more interesting, some examples seem to evoke both a literal and a non-literal reading ('b'), often in a deliberate play with words, as with the expression *hold the baby* in (13). Finally, we recently introduced a seventh label 'embedded' ('e') for cases in which the target expression is embedded in a larger figurative context. For instance in (14), the expression *fly high* is used in a context where the addressee is metaphorically compared to a bird (14). Another example is given in (15). While meta-linguistic, mixed, and embedded usages are rare (see Figure 4), we believe that they may also be particularly interesting for linguistic studies (especially mixed and embedded usages) and thus decided to annotate them as separate categories.

(12) It has long been recognised that expressions such as to pull someone's leg, to have a bee in one's bonnet, to kick the bucket, to cook someone's goose, to be off one's rocker, round the bend, up the creek, etc. are semantically peculiar. [BNC FAC]

(13) Left holding the baby, single mothers find it hard to fend for themselves. [BNC CRA]

(14) You're like a restless bird in a cage. When you get out of the cage, you'll fly very high. [BNC FR6]

(15) Political prudence and the dangers of a frontal attack on the Church restrained them to the sale of common lands and the abolition of civil entails, 'pulling up by the roots the tree which bears such bitter fruits'. [BNC FB7]

During the annotation process it became also clear that we had to deal with expressions which have more than one non-literal reading. An example is *cover the ground*, which can mean (i) 'deal (thoroughly) with a subject' as in (16) as well as (ii) 'travel or pass a certain distance or area' as in (17). Additionally, it also has a literal reading of 'spreading over a certain area' as in (18).

(16) The ground covered by both books is, in the early stages, fairly similar. [BNC B77]

(17) Only 5ft 8ins tall and under 10st in weight, he covers the ground quicker than anyone I have recently seen, with the exception of one he resembles, South Africa's Jonty Rhodes. [BNC BN9]

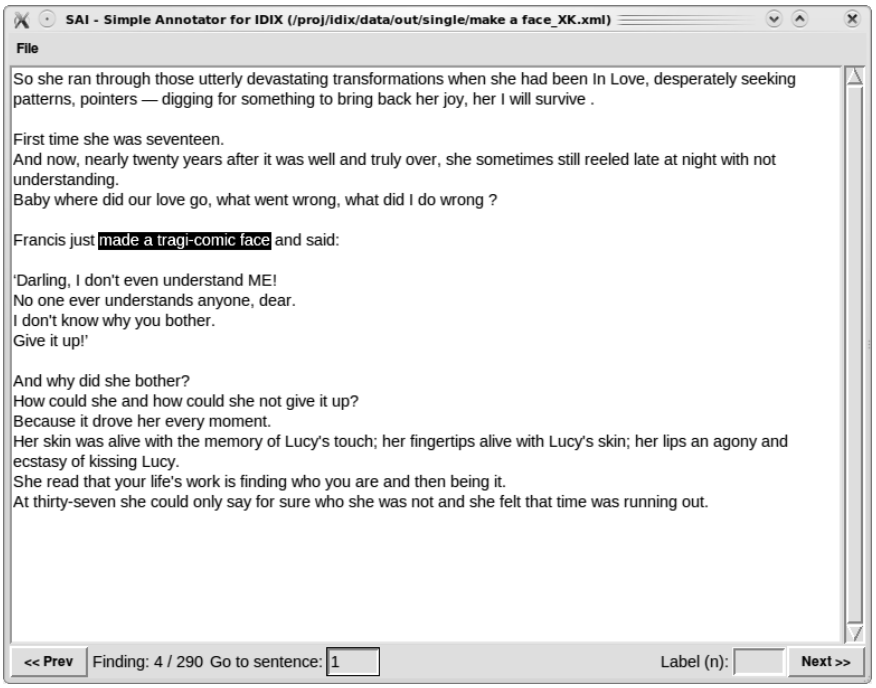

Figure 3: SAI tool with highlighted instance of the target expression *make a face*

(18) The most reliable standby for climbing or simply to cover the ground, has to be ivy. [BNC C9C]

We distinguish these meanings by assigning an id-number to each non-literal sense and then suffixing this id-number to the label in the annotation. A documentation file providing information on the number of non-literal senses and a short definition for each will be made available as part of IDIX.

2.5. Corpus Statistics and Evaluation

Work on the IDIX corpus is still ongoing. We aim to expand the corpus to at least 100 idioms. So far we have annotated 78 expressions. This amounts to 5,836 instances in total (excluding false extractions), i.e., on average around 75 instances per expression, ranging from one instance for *lower the bar* to 540 for *ring the/a bell*. Table 1 gives an overview of the 20 most frequent expressions and their distributions of annotation labels. It can be seen that expressions behave quite differently with respect to the frequency of literal an non-literal usages. For some expressions, like *make a face*, the non-literal usage is much more frequent. For others, like *show one's teeth* or *pull the trigger*, literal usage is much more frequent.

Figure 4 shows the proportions of annotation labels in the currently annotated data set. It can be seen that a small majority of 49.4% was annotated as 'literal', 45.61% of the instances were annotated as 'non-literal', 0.15% as 'meta-linguistic', 0.69 % as 'both literal and non-literal readings' and the rest as 'undecided'. Cases where an expressions is embedded in a larger figurative context have not yet been annotated separately; they are currently included in the set of idiomatic usages but will be annotated as 'embedded' in the release version of the corpus. However, we believe that these cases are also relatively rare. The relatively large proportion of literal usages can probably be attributed to the fact that we extract both canonical forms, for which idiomatic usages tend to be more frequent, and non-canonical forms, for which literal usages occur more frequently (Riehemann, 2001).

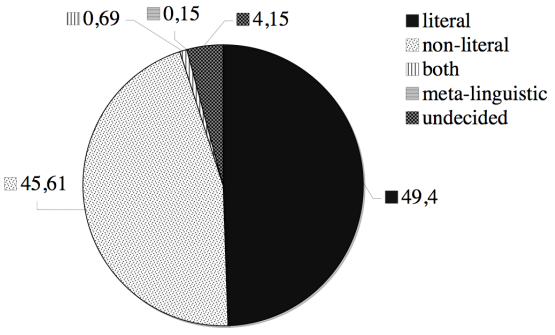

Figure 4: Percentage of labels in the currently annotated data set

To assess annotation quality, 24 idioms (1,136 examples) were annotated independently by two annotators. The overall inter-annotator agreement was 93.19% . We also computed the Kappa statistic (Krippendorff, 1980), which corrects the percentage agreement for

idiom	instances	lit. (l)		non-lit. (n)		mixed (b)		meta-ling. (m)		unclear (?)	
ring the bell	540	389	(72)	137	(25)	0	(0)	0	(0)	14	(3)
raise one's eyebrows	468	405	(86)	54	(12)	0	(0)	0	(0)	9	(2)
draw the line	427	120	(28)	266	(62)	0	(0)	0	(0)	41	(10)
pay dividends	354	226	(64)	128	(36)	0	(0)	0	(0)	0	(0)
hold somebody's hand	340	289	(85)	29	(9)	12	(3)	0	(0)	10	(3)
make a face	203	7	(4)	173	(85)	0	(0)	0	(0)	23	(11)
deliver the goods	183	106	(58)	69	(38)	7	(3.5)	1	(0.5)	0	(0)
get the message	177	62	(35)	101	(57)	1	(0.5)	0	(0)	13	(7.5)
carry weight	173	29	(17)	127	(73)	0	(0)	0	(0)	17	(10)
lick one's lips	159	150	(94)	7	(4.5)	0	(0)	0	(0)	2	(1.5)
reach the top	156	121	(78)	33	(21)	0	(0)	0	(0)	2	(1)
answer the call (of duty)	123	50	(40)	49	(40)	0	(0)	0	(0)	24	(20)
bear fruit	119	19	(13)	99	(83)	0	(0)	0	(0)	1	(1)
strike a chord	114	6	(5)	106	(93)	0	(0)	0	(0)	2	(2)
gain ground	108	12	(11)	95	(88)	1	(1)	0	(0)	0	(0)
do one's homework	104	36	(35)	65	(62)	0	(0)	0	(0)	3	(3)
cover the ground	102	23	(22.5)	73	(71.5)	1	(1)	1	(1)	4	(4)
show one's teeth	87	82	(94)	3	(3.5)	0	(0)	0	(0)	2	(2.5)
cut someone's throat	87	65	(75)	14	(16)	0	(0)	0	(0)	8	(9)
pull the trigger	84	77	(92)	5	(6)	1	(1)	1	(1)	0	(0)

Table 1: The 20 expressions with the most instances, excluding erroneous extractions; numbers per label and idiom (percentage per label/idiom)

expected chance agreement. A Kappa score of $K \geq .80$ is considered as reliable agreement. The Kappa score on our data set is $K = .87$. Moreover, only a minority of disagreements (37.66% of the disagreements, 2.56% of the instances overall) involved cases where one annotator had classified an instance as literal and the other as idiomatic. Disagreements about 'literal' or 'non-literal' vs. 'mixed/undecided' accounted for another 35% of all disagreements. Hence, overall, the annotators agreed rather well, especially with respect to the crucial distinction between literal and non-literal usages.

3. Conclusion

In this paper, we introduced the IDIX corpus, which provides information about the usage of potentially idiomatic expressions in a given discourse context. We distinguish six different usages: 'literal', 'non-literal', 'mixed', 'embedded in larger figurative context', 'meta-linguistic', and 'unclear/undecidable'. In addition, we also distinguished between different non-literal senses as far as they exist for a given target expression. We found that the distinction between different usages can be made relatively reliably. In our data, literal readings are most common, accounting for just under half of all cases. Non-literal readings account for most of the remaining cases, while other usages are rarer.

The IDIX corpus will be made available as an add-on for the BNC.[6] Like the BNC, IDIX covers data from several domains. Furthermore, for each of the chosen expressions, the annotation is exhaustive, in the sense that all BNC occurrence of the target expression are annotated. We believe that this resource will be useful for NLP researchers working on context-dependent, token-based idiom detection, as well as linguists working on corpus-based studies of non-literal language.

The annotation effort is still ongoing. For the first release, we aim to annotate at least 100 expressions. While IDIX provides annotations for all occurrences of a given expression, i.e., the coverage is complete with respect to the expression, we did not aim for complete coverage of individual texts, i.e., the texts may contain additional idioms or other figurative usages which are not annotated. For certain linguistic or computational linguistic studies it would also be interesting to exhaustively annotate sample text with all figuratively used expressions. We plan to complement IDIX with such a resource in the future. Exhaustive annotation of texts is a harder task, especially if the

[6]The corpus will be downloadable soon from `http://www.coli.uni-saarland.de/projects/comodis/idix.html`.

annotation is not restricted to idioms but extended to metaphor, metonymy and other cases of figurative language. We are currently working on annotation guidelines for this task.

Acknowledgements

This work was funded by the German Research Foundation DFG within the Cluster of Excellence Multimodal Computing and Interaction (MMCI).

4. References

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

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

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL-07 Workshop on A Broader Perspective on Multiword Expressions*.

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

Julia Birke. 2005. A clustering approach for the unsupervised recognition of nonliteral language. Master's thesis, School of Computing Science. Simon Fraser University.

Francis Bond and Satoshi Shirai. 1997. Practical and efficient organization of a large valency dictionary. In *Workshop on Multilingual Information Processing Natural Language Processing Pacific Rim Symposium*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastia Padó, and Manfred Pinkal. 2006. The SALSA corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC-06*.

Paul Cook, 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*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

A.P. Cowie, R. Mackin, and I.R. McCaig. 1997. *Oxford dictionary of English idioms*. Oxford University Press.

Mona Diab and Madhav Krishna. 2009a. Handling sparsity for verb noun MWE token classification. In *Proceedings of the EACL Workshop on Geometrical Models of Natural Language Semantics*, pages 96–103.

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

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL-06*.

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

Fabienne Fritzinger, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb tiples on token level. In *Proceedings of LREC-10*.

Laurie Gerber and Jin Yang. 1997. Systran MT dictionary development. In *Proceedings of the Fifth Machine Translation Summit*.

Chikara Hashimoto 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.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006a. Detecting Japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40(3-4):243–252.

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

Graham Katz 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*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

David D. Lewis and W. Bruce Croft. 1990. Term clustering of syntactic phrase. In *Proceedings of SIGIR-90, 13th ACM International Conference on Research and Development in Information Retrieval*.

Linlin Li and Caroline Sporleder. 2009a. Classifier combination for contextual idiom classification without labelled data. In *Proceedings of EMNLP-09*.

Linlin Li and Caroline Sporleder. 2009b. A cohesion graph based approach for unsupervised recognition of literal and nonliteral use of multiword expressions. In *Proceedings of the ACL 2009 Workshop on TextGraphs-4: Graph-based Methods for Natural Language Processing*.

Linlin Li and Caroline Sporleder. 2010. Using Gaussian Mixture Models to detect figurative language in context. In *Proceedings NAACL-2010 Short Papers*.

Dekang Lin. 1998. Using collocation statistics in information extraction. In *Proceedings of MUC-7*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324.

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

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