Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

11 | 2012
Varia

So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text

Alexis Palmer, Caroline Sporleder and Linlin Li

Electronic version
URL: http://journals.openedition.org/discours/8731
DOI: 10.4000/discours.8731
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Alexis Palmer, Caroline Sporleder and Linlin Li, « So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text », *Discours* [Online], 11 | 2012, Online since 23 December 2012, connection on 30 April 2019. URL : http://journals.openedition.org/discours/8731 ; DOI : 10.4000/discours.8731

discours
Revue de linguistique, psycholinguistique et informatique
http://discours.revues.org/

So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text

Alexis Palmer
Saarland University

Caroline Sporleder
Saarland University

Linlin Li
Saarland University

Alexis Palmer, Caroline Sporleder et Linlin Li, « So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text », *Discours* [En ligne], 11 | 2012, mis en ligne le 23 décembre 2012.

URL : http://discours.revues.org/8731

Titre du numéro : *Varia*
Coordination : Olivier Ferret et Nicolas Hernandez

revues.org
CENTRE POUR L'ÉDITION ÉLECTRONIQUE OUVERTE
CENTRE FOR OPEN ELECTRONIC PUBLISHING

discours

Presses universitaires de Caen

So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text

Alexis Palmer
Saarland University

Caroline Sporleder
Saarland University

Linlin Li
Saarland University

Lexical cohesion is an important device for signaling text organization. In this paper, we investigate to what extent a particular class of expressions which can have a non-literal interpretation participates in the cohesive structure of a text. Specifically, we look at five expressions headed by a verb which – depending on the context – can have either a literal or a non-literal meaning: *bounce off the wall* ("to be excited and full of nervous energy"), *get one's feet wet* ("to start a new activity or job"), *rock the boat* ("to disturb the balance or routine of a situation"), *break the ice* ("to start to get to know people, to overcome initial shyness"), and *play with fire* ("to take part in a dangerous or risky undertaking"). We look at the problem both from an empirical and a computational perspective. The results from our empirical study suggest that both literal and non-literal expressions exhibit cohesion with their textual context, but that the latter appear to do so to a lesser extent. We also show that an automatically computable semantic relatedness measure based on search engine page counts correlates well with human intuitions about the cohesive structure of a text and can therefore be used to determine the cohesive structure of a text automatically with a reasonable degree of accuracy. This investigation is undertaken from the perspective of computational linguistics. We aim both to model this cohesion computationally and to support our approach to computational modeling with empirical data.

Keywords: idioms, literal usage, non-literal usage, multi-word expressions, cohesion, semantic relatedness, lexical chains, natural language processing, annotation

1. Introduction

1 Computational linguistics deals with the automatic processing and analysis of natural language texts or speech. There are two main branches: one branch views automatic language processing as a supplementing technology for linguistics. The aim is to develop and formalize automatable methods for language analysis that can then be used to verify or falsify linguistic theories. The second branch deals with language processing from an engineering perspective. Here the aim is to automate language

processing (both with respect to analysis and with respect to generation) for real-life applications, such as machine translation, automatic extraction of information from text, or automatic text rewriting (e.g., text summarization, text simplification). The work reported in this paper is situated in the engineering-oriented branch of computational linguistics but also aims to provide an empirical underpinning for some of the assumptions made by one particular approach to a particular problem in automatic language processing. The problem we are concerned with is the automatic detection of non-literally used multi-word expressions (MWEs) in running text, such as *break the ice*. Given a computer-readable dictionary of MWEs this problem is relatively trivial for those expressions that cannot have a literal reading (i.e., that are always used non-literally). An example is *shoot the breeze* (meaning "to engage in idle conversation", "to chat"). It is very difficult to come up with a real-life context in which this expression would refer to a "shooting" event that involves "a breeze". However, a fairly large group of MWEs can have a literal interpretation as well and these MWEs are thus potentially ambiguous between literal and non-literal usage (i.e., ambiguous for a computer; humans typically do not have any difficulties picking the right interpretation given the context). This ambiguity arises particularly frequently with MWEs that consist of a verb plus one or more complements (e.g., V+NP, V+PP), such as *break the ice* or *swim against the tide*. These expressions have to be disambiguated by looking at the context in which they occur. This is easy for humans but difficult for a computer. Humans process language more or less incrementally. They hear or read part of a sentence, process it, and make a hypothesis about how it will be continued. Natural language processing software, on the other hand, typically computes meaning representations in a non-incremental, bottom-up fashion by applying several processing steps in sequence, starting with automatic part-of-speech assignment, continuing with syntactic parsing, word sense disambiguation, semantic parsing and finally discourse parsing. Ambiguous expressions such as *break the ice* pose a problem for this setup because whether an expression is used non-literally (i.e., is an MWE) ideally has to be known fairly early on in the process since it potentially affects syntactic and semantic parsing. For instance, syntactic and semantic parsing both typically exploit statistics about selectional preferences which depend on whether the expression in question is a non-literally used MWE or its completely compositional, literal counterpart. Indeed it has been shown that a significant number of errors made by syntactic parsers can be attributed to (a failure to recognize) MWEs (Baldwin et al., 2004). Hence, it is necessary to identify non-literal expressions before the meaning of the sentence is computed. This requires a relatively knowledge-poor approach that does not rely on (full) syntactic or semantic processing of the context but rather utilizes statistics that distinguish literal and non-literal usages. This is an active research area in computational linguistics and it is this problem that we are concerned with in this paper.

2 One approach that has been suggested for addressing this problem is based on lexical cohesion (Sporleder & Li, 2009). This work employs a statistical model of

semantic relatedness between words to compute the overall cohesive structure of a text. The model then classifies the target expression, depending on whether a literal or a non-literal interpretation fits better with the overall cohesive structure.

3 However, the detection of non-literal expressions is not the only area of computational linguistics that has made use of the concept of "cohesion". Computational applications that make use of cohesion range from the detection of malapropisms (Hirst & St-Onge, 1998) over word sense disambiguation (Okumura & Honda, 1994) and topic segmentation (Hearst, 1997) to automatic text summarization (Barzilay & Elhadad, 1997). The reason why cohesion is such a useful concept in computational linguistics is that cohesion is fairly easy to compute in a knowledge-poor fashion, that is relying on surface cues and statistics rather than on deep linguistic processing, and that it provides a great deal of information about the internal structure of a text.

4 Broadly, the term "cohesion" refers to the manner in which words or syntactic features connect individual sentences and clauses to their discourse context. Halliday and Hasan (1976) propose five classes of cohesion: conjunction, reference, substitution, ellipsis, and lexical cohesion. Lexical cohesion covers various semantic relationships between the lexical items (primarily words and MWEs) in a text, ranging from literal repetition (called "reiteration" by Halliday and Hasan) to weaker semantic relationships (so-called "non-classical relations" [Morris & Hirst, 2004]), such as that between *wet* and *bathtub*, or that between *laugh* and *joke*.

5 Lexical cohesion is especially interesting for studying textual organization because, in addition to being the most frequent class of cohesive ties (Hoey, 1991), it tends to be a global phenomenon. In other words, entire texts can be analyzed in terms of *chains* of lexically cohesive words, which may span large segments of the text or even the text as a whole, if the chain refers to the central topic of a discourse. Lexical ties thus serve as one indication of the overall structure and organization of a text, for example, with respect to the main topics addressed by the text and the distribution of those topics throughout the text.

6 In this paper, we address a particular aspect of lexical cohesion, namely how non-literally used MWEs fit into the cohesive structure of a text. Given that non-literal MWEs are semantically more or less opaque and non-compositional, how might we expect such expressions to fit into the overall structure of the text? Consider, for example, the literal and non-literal meanings of the string *spill the beans*. When literally used, we would expect to find the expression in a semantic context having something to do with food, dining, cooking, or perhaps preparing or storing foodstuffs. Alternatively, the expression may be part of a text with a clumsy or accident-prone participant. If the expression is used instead with its non-literal meaning, an entirely different semantic domain would be expected, most likely having to do with keeping or revealing secrets.

7 We are interested in (i) whether it is possible to find cohesive ties between a non-literally used expression such as *spill the beans* and the surrounding context,

(ii) whether such ties are stronger or weaker than for the component words of the literal counterpart of the expression, and (iii) how such cohesive links can be modeled computationally. We evaluate whether the cohesive links found automatically are identical or at least similar to those annotated by humans. We also explore whether deviations between the two are due to errors made by the automatic method or whether humans pick up on a different type of cohesion than is captured by the automatic tool. Our work thus combines empirical and computational approaches. With this study, we add to a significant body of prior work on lexical cohesion, both in the linguistics (Hoey, 1991; Tanskanen, 2006) and the computational linguistics communities (Okumura & Honda, 1994; Barzilay & Elhadad, 1997; Hearst, 1997; Hirst & St-Onge, 1998).

2. Human evaluation of cohesive chains

8 In this study, we aim to better understand the interaction between lexical cohesion chains and a particular type of expression which can be used either literally or non-literally. First, we carry out a small-scale annotation study in which the goal of annotation is to identify and label the lexical cohesion chains in which instances of these expressions (and their individual component words) participate. These annotations are then used to evaluate the strength of cohesion with both the literal meaning of the expression and its metaphorical meaning. The annotation methodology is discussed in 2.1; here we describe our texts.

9 For human expert identification of cohesive links, we carried out a small-scale annotation study, using texts from Sporleder and Li's (2009) data set which was itself extracted from a large newswire text corpus (the *Gigaword* corpus) [1]. Sporleder and Li's data set consists of texts containing expressions which can be used literally as well as non-literally. From these, we chose five expressions to work with: *bounce off the wall* (henceforth: **wall**), *get one's feet wet* (**feet**), *rock the boat* (**boat**), *break the ice* (**ice**), and *play with fire* (**fire**). The particular expressions were selected in part based on how accurately their instances are classified by Sporleder and Li's automatic method for distinguishing literal and non-literal usage. For *bounce off the wall*, Sporleder and Li's method erroneously classifies many literally used examples as non-literal, and the reverse is true for *get one's feet wet*. *Rock the boat*, on the other hand, was included because the performance of the classifier is relatively high for this expression; *break the ice* and *play with fire* were selected more or less randomly. Table 1 shows the accuracies obtained by Sporleder and Li's cohesion-based classifier for each of the five expressions included in the present study.

1. The newswire genre was chosen because (i) a large amount of data is available in electronic form for this genre, and (ii) most natural language processing applications focus on newswire, partly due to its availability and partly due to the fact that for this text type there is a high demand for language processing applications such as information extraction and text summarization. This distinguishes newswire from other text types such as fiction.

Expression	Detection accuracy
bounce off the wall	47.82%
get one's feet wet	64.33%
rock the boat	98.95%
break the ice	85.03%
play with fire	82.33%

Table 1. Idiom detection accuracies for Sporleder and Li's (2009) cohesion-based classifier

10 For each expression, we randomly chose four texts from the data set for annotation: two with literal uses and two with non-literal uses. The texts were annotated by the three authors of this paper, who all have a background in either linguistics or computational linguistics and who are either native or near-native speakers of English. Two annotators labeled the texts in their entirety, and the third annotator labeled the portion of text immediately surrounding the expression of interest, with a window of approximately two paragraphs in each direction. Each annotator identified and labeled two chains (i.e., two sets of semantically-related words) for each text: one for the literal meaning of the target expression (henceforth referred to as the **literal chain**) and one for the non-literal meaning (henceforth referred to as the **non-literal chain**).

11 Two hypotheses are under investigation in this study. The first is that literal and non-literal meanings of an expression can be distinguished on the basis of lexical chains. Our expectation is that one chain – the one for the meaning intended by the author – should always be noticeably stronger than the other. Cohesion with the non-intended meaning should be merely accidental, and one might expect that the authors try to deliberately minimize it to avoid confusion[2]. This expectation is intuitively obvious and perhaps uninteresting, but automatically disambiguating such expressions is far from trivial, and lexical chains are a potential signal for an automatic system. If the expectation is borne out by empirical evidence, the relative strength of the two chains in a text can be used for disambiguation.

12 Our second hypothesis is that non-literal usages tend to exhibit weaker cohesion (with the non-literal chain) than literal usages do (with the literal chain). If this is indeed the case it will be more difficult to find strong evidence for non-literal usage based on cohesion alone. This should be taken into account by automatic systems; for example, it might be useful to have a lower prediction threshold for predicting non-literal usage, thereby predicting non-literal usage even if the cohesive evidence for this is only moderate. This would also be justified by the fact that the

2. Sometimes one can observe intended cohesion with both meanings, usually due to a deliberate play on words.

prior probability of non-literal usage is higher than that of literal usage for most of the expressions we investigate, i.e., non-literal usage is more frequent. This is certainly true for the news domain from which our data stems (and has been shown empirically by Sporleder and Li [2009]).

2.1. Annotation process and decisions

13 The basic annotation task here is to identify and mark each word or MWE belonging to each of two cohesion chains in a document. More precisely, for each occurrence of a potentially non-literal expression, two cohesion chains are identified and annotated. It happens that in our corpus each document contains just one expression of interest. Any individual word may participate in both chains, though such cases are rare. Of over 600 lexical items marked (across the 20 texts), only 10 were marked as participating in both the literal and the non-literal chain. For example, the word *warmth* belongs weakly to the non-literal chain for a usage of *break the ice* (via the paraphrased meaning, see Table 6) and also, through antonymy, to the literal chain for the component word *ice*.

14 Annotating cohesive chains is a notoriously difficult task, since it is often a matter of debate whether, to what degree, and in what way two words are semantically related. Relatively few empirical studies have looked into human intuitions regarding lexical cohesion, and those that have done so generally report low inter-annotator agreement (Hollingsworth & Teufel, 2005; Beigman Klebanov & Shamir, 2006; Morris & Hirst, 2006; Stührenberg et al., 2007; Cramer et al., 2008). To alleviate this problem to some extent, following preliminary annotation of seven texts, the annotators discussed potential problems, arriving at general guidelines for the task. Each annotator then adjusted their preliminary annotations and independently labeled the remaining thirteen texts. We show annotator agreement for the texts discussed and not discussed in Table 5, but all twenty texts are treated the same in the discussion of results. The annotation guidelines that emerged are discussed below as points 2.1.1 to 2.1.6.

2.1.1. Literal chains

15 Annotators identified literal and non-literal chains for all texts, regardless of whether the target expression itself is used with its literal or its non-literal meaning. Two anchor words were identified for each idiom, corresponding to the semantically most contentful words of the expression, e.g., a verb and a noun in V + NP or V + PP constructions. Annotators marked literal cohesion chains for both anchor words. The idea of anchors is reminiscent of Beigman Klebanov and Shamir (2005), who instruct their annotators to identify anchors for concepts in a text, but we differ in that we predefine the anchors, as we are interested only in specific chains, i.e., those related to the literal and the non-literal meaning of the (elements of the) target phrase. The anchor words and the number of links to each appear in Table 2. In all but one case, the second anchor word, typically a noun, receives many more cohesive links than the first. This confirms an intuition that nouns

exhibit more cohesion with their context or at least participate in more easily identifiable cohesion relations than verbs.

1st anchor		**2nd anchor**	
bounce	**15**	wall	6
rock	5	boat	**70**
break	13	ice	**36**
feet	20	wet	**89**
play	3	fire	**38**

Table 2. Literal chains: Number of cohesive links to anchor words of target expressions

2.1.2. Non-literal chains

16 For the non-literal chains, annotators marked words exhibiting lexical cohesion with the non-literal meaning of the target expression. Because that meaning can be difficult to pin down, we developed a set of paraphrases for each idiom. These paraphrases were used both to guide human annotation and for automatic computation of cohesion (see Section 3, Table 6).

2.1.3. Semantic relationships

17 After some discussion, the decision was made to mark only shallow, lexically-based semantic relationships between words. Cohesive links based on world knowledge (for example, linking *pasta* with *marathon* via knowledge of the practice of carb-loading) were not marked. By excluding such links we aimed to make the annotation more objective and reliable. Unlike some other annotation studies (Stührenberg et al., 2007; Cramer et al., 2008), we did not prespecify or restrict the types of semantic relationships that could be marked. In particular, annotators were asked to mark not only classical relations such as synonymy, antonymy, homonymy, and meronymy, but also non-classical relations. Furthermore, we did not specify a distance cut-off for cohesive links; all links within the context provided to the annotators were included.

18 Rather than marking particular types of semantic relations, we distinguished only between two types of cohesive links: weak and strong. Strong links were annotated for strong semantic relationships, such as that between *wet* and *water*. Weak links were annotated for more indirect relationships, e.g., between *wet* and *diving*, which are related via the concept of *water*. While strong links are relatively easy to identify, weak links often require some degree of inference; annotators tend to disagree more about these. It should also be noted that the distinction between strong and weak links is not totally clear-cut but rather is open to interpretation. For illustration, Table 3 lists some strong and weak links for the word *wet*. Note that antonyms were also included as (strong) semantic links.

Strong	shower, rain, bath, river, water, dry, drizzle, ocean
Weak	boat, marina, dolphins, fish, dockside, island, sandbar

Table 3. Strong and weak links for *wet*

19 Since we ask annotators to find relations between context words and a target concept (rather than between linearly ordered context words), the cohesive structures we identify are not linear structures but rather resemble clusters or graph-structures. This is in line with several previous studies which found that humans find it difficult to identify linear structures and prefer net-like structures (Stührenberg et al., 2007; Cramer et al., 2008). For convenience, we will continue to use the term (cohesive) "chain" throughout this paper.

2.1.4. Words vs. concepts

20 Previous annotation studies differ somewhat in whether they annotate semantic relations between words (i.e., tokens) or concepts (i.e., types). For example, Beigman Klebanov and Shamir (2005) annotate concepts; each lemma is only considered once, and it is assumed that repeated occurrences of a lemma will all link back to its first mention. Hollingsworth and Teufel (2005), on the other hand, annotate word tokens. In this study, we also annotate word tokens rather than concepts, i.e., repeated occurrences of a word are considered separately, and a single chain may consequently contain multiple occurrences of a given lemma. Theoretically, different occurrences of a given lemma could also participate in different chains, depending on the context in which they appear. However, in practice this case did not arise. The choice to annotate word tokens rather than concepts was motivated by the fact that we want to measure chain strength, and we define chain strength (partly) in terms of the number of elements contained in a chain, adopting the assumption that repetitions strengthen a chain. An individual word token was also allowed to participate in several chains at once, e.g., in the literal and the non-literal chain, though in practice this also happened rarely.

2.1.5. Markables

21 Many studies on lexical chaining assume that only nouns can participate in chains (Morris & Hirst, 2006; Stührenberg et al., 2007; Cramer et al., 2008), though some also include other content words (Hollingsworth & Teufel, 2005; Beigman Klebanov & Shamir, 2006). Adjectives and verbs, in particular, have been found to participate in lexical chains, albeit less frequently than common nouns, while proper nouns and function words participate very little in the lexical cohesive structure of a text (Beigman Klebanov & Shamir, 2006). In our case, annotators were asked to mark all content words (including MWEs, see below) which exhibit cohesive links with the target concept. Named entities were left unmarked, because relating these semantically to the context also typically requires world knowledge. For example, it can be argued that *Wayne Rooney* is semantically related to *ball* but making this connection requires world knowledge, i.e., one has to know that Wayne Rooney is a football player.

2.1.6. MWEs

22 The human annotators marked relevant MWEs as participating in cohesive links, with each expression representing a single link. However, MWEs pose a particular challenge for automated text-processing systems, and the method we use to compute lexical cohesion does not accommodate MWEs. This has the result that some prominent cohesive links are ignored in the automated processing. Example [1] is from a text with a non-literal occurrence of *get one's feet wet*. The text is a report on the small but growing number of women in talk radio and the obstacles they face on that career path.

[1] That's not due to gender bias, although **breaking into the field** is harder for a woman, McCoy said. "I think it might be tougher for a woman to **get started** than a man".

23 Both *break into the field* and *get started* are reasonable (though not perfect) paraphrases for the idiom and as such form strong cohesive links. Of the individual words in the two phrases, only one of each link (*field* and *started*) independently exhibits lexical cohesion with the non-literal meaning.

24 In other cases, though, each content word of a MWE exhibits cohesion with the target expression. In those cases, the links are preserved by marking each word separately. [2], from the same text as [1], shows two such cases (*enter the field* and *developing skills*).

[2] That is changing, though, as more women **enter** the **field**... Now that more are, they are *getting their feet wet* and **developing skills**.

25 The annotation was done in three passes. In the first pass, the annotators only read the text without marking any cohesive elements to obtain a general idea of the topic and content of the text. In the second pass, only the most obvious cohesive links were marked (typically the strong links), while in the third pass the annotators looked for further links (typically weak links).

26 For a given expression, annotators first labeled texts with non-literal usages and then turned to those with literal usages. Texts were annotated in this order so that annotators could increase their familiarity with the semantic content of the target expression, as well as with its potential cohesive links, before tackling the more difficult case of annotating non-literal chains for literal usages. This decision was made under the hypothesis that non-literal chains are more prominent for non-literal usages of the target expression.

2.1.7. Gold standard

27 The annotations described above also need to serve as a "gold standard" (following standard terminology in computational linguistics) against which to evaluate the automatically-produced analysis described later in the paper. The gold standard (GS)

needs to provide just a single label per word in the text, thus requiring adjudication over the three sets of human annotations. Where annotators disagreed on the status of a given word with respect to the cohesion chains, in most cases we simply took a majority vote to choose the GS annotation. Two different situations required something other than a majority vote. The first, most obvious, situation involves the portions of texts marked by two rather than three annotators. When the two annotators disagreed, the GS annotation was determined based on the linguistic intuition of the adjudicator. The second situation occurred primarily when one or more annotator showed a lack of consistency in their labeling of a given word within a given text. For example, in one text all three annotators marked an early occurrence of *diver* as belonging to the literal chain for the expression *get one's feet wet*. A later occurrence of *diver* was marked as belonging to the chain by only one annotator and given no marking by the other two. Because the annotations for individual lexical items are relevant at the type rather than the token level (i.e., the semantic relation holds between lemmas, or rather senses, rather than between word occurrences), such cases were treated as oversights and marked in the GS as belonging to the relevant chain, even though such marking, strictly speaking, goes against the majority vote of the annotators.

2.2. Findings

28 The raw results of annotation are shown in Table 4. For each text, the table shows the number of lexical items in the two cohesion chains for the text following adjudication to a GS. Figure 1 presents one full text and its complete (adjudicated) annotation.

29 To determine the reliability of our annotation, we computed the correlation between the first two annotators using Pearson's product-moment correlation, as implemented in the R statistical software package [3]. The top half of Table 5 shows aggregate correlation figures for all texts, broken down between those with literal uses of the target expressions and those with non-literal uses. The bottom half of the same table distinguishes texts which were discussed by the annotators from those which were annotated entirely independently.

30 It can be seen that the correlation is generally good, even for the texts that were not discussed. Overall, the correlation is higher for literal than for non-literal chains. Hence it seems that it is easier to agree on related words for literal usages, while non-literal usages are fuzzier and therefore less easy to annotate.

31 We also computed correlation with respect to link strength (i.e., strong vs. weak) for annotators one and two. Again using the R implementation of Pearson's product-moment correlation, for all pairs marked by both annotators as participating in a cohesive chain, correlation was measured as 0.4639.

3. The R package is available from http://www.r-project.org/.

Expr-textid	Type	Literal	Non-literal
bounce-6	non	0	**3**
bounce-12	lit	**9**	5
bounce-43	non	5	**10**
bounce-48	lit	**7**	12
boat-6	non	1	**29**
boat-125	lit	**17**	13
boat-233	non	1	**24**
boat-420	lit	**56**	3
ice-49	lit	**27**	2
ice-149	non	6	**16**
ice-347	non	5	**12**
ice-464	lit	**11**	5
feet-37	non	2	**17**
feet-114	lit	**42**	10
feet-165	non	21	**6**
feet-169	lit	**44**	15
fire-200	non	16	**10**
fire-304	non	2	**8**
fire-581	lit	**14**	2
fire-589	lit	**8**	0

Table 4. Cohesion chains marked in all texts

Texts	Literal chains	Non-literal chains
ALL	0.8115	0.7354
Literal usage	0.8189	0.6724
Non-literal usage	0.8031	0.7747
Literal discussed	0.8142	0.6639
Literal not discussed	0.8235	0.7061
Non-literal discussed	0.9594	0.9388
Non-literal not discussed	0.6859	0.6763

Table 5. Correlations between annotators 1 and 2

BOLD items belong to literal chain
BOLD$_x$ subscript x indicates anchor word (1st or 2nd content word of target expression)
ITALIC items belong to non-literal chain

DIVER$_2$ TAKES WORKADAY PERILS IN **STRIDE**$_1$
SAN ANTONIO (BC-PRO-DIVER-HNS)

Rather than dressing for success, Jeff Davila dresses for descent _ **diving**$_2$ mask, **wet**$_2$ suit, gloves, **boots**$_1$, knife and a 30-pound weight belt.

In his short ***career*** as a commercial **diver**$_2$ at Walt Disney World, Davila had his share of close encounters and shaky moments. Davila, 30, recalls cleaning the bottom of a 15-foot-deep **lake**$_2$ at Disney's Animal Kingdom in Orlando, Fla., under the watchful gaze of a 6-foot **alligator**$_2$. At one point, the two came face to face.

"It didn't attack, but it was very unnerving," he said. "It was about the longest four hours of my life."

The San Antonio native, who waited tables and played in rock bands before, literally, ***taking the plunge***, says he has always been a city dweller with a fascination for the **ocean**$_2$. A couple of years ago he enrolled in an eight-month, $12,000 program at Ocean Corp., a **diving**$_2$ school in Houston. He had no previous **diving**$_2$ ***experience***.

"When I was a kid, I was into Jacques Cousteau, National Geographic and music," he said recently. "I've loved the **ocean**$_2$ ever since I could **walk**$_1$. My father would just take me out there and **dip**$_2$ my **feet**$_1$ in the **water**$_2$. We'd go to Corpus (Christi, Texas) every summer."

He has been **diving**$_2$ commercially for only about a year, but recently left Disney, where he was overqualified.

"Basically, if you work for Disney, you're a glamorized janitor and animal feeder," said John Wood, president of Ocean Corp.

Davila said he worked for Disney to get his feet wet, so to speak. Now he is looking for a job with an **underwater**$_2$ welding company. A professional **diver**$_2$ can earn up to $100,000 a year.

Wood said the ***training*** to become a professional **diver**$_2$ is tough because the nature of the business is unpredictable.

"We'll scare the bejesus out of you," he said. "That's our job. With ***training***, you can ***train*** away the panic response from a person. We have a protocol to deal with accidents. If something goes wrong, they have a procedure and a technique they follow."

After years of waiting tables, playing music and a sporadic college education, Davila has found his niche in life. He loves the work, the lifestyle and challenges _ challenges such as conducting inspections in zero-visibility **water**$_2$.

"It's a weird feeling," he said. "You feel like you're being watched. Imagine going into a closet, closing the door, and just working on something for four hours."

Figure 1. Adjudicated human annotations for one complete text

32 As expected, agreement on weak links was generally lower than on strong links. For example, our annotators disagreed about whether *island*, *diving* and *sailing* form weak or strong links with *wet*[4]. Similarly for *lifeguard*, *lobster*, and *diving mask* there was disagreement as to whether they formed a weak link with *wet* or no link at all[5]. Cases in which a word was classified as strong by one annotator while the other annotator decided there was no link were rare and many of these were either genuine annotation mistakes, i.e., oversights by one annotator, or they involved links between words with different parts-of-speech. For instance, one annotator identified a strong link between *plunge* and *wet* while the other annotator decided there was no link. This is to be expected since identifying semantic relations across parts-of-speech is notoriously more difficult than within the same part-of-speech.

33 Once the human annotations had been adjudicated and a GS produced, we computed the strengths of the annotated chains. It is common in computational linguistics to model the strength of a chain in terms of its length, i.e., the more word tokens a chain contains, the stronger it is. We adopted this measure here, first making no distinction between strong and weak links, and later recalculating by giving strong links twice the weight of weak links. There was no quantitative change in the results from taking into account the two types of links. As expected, the chains for the intended meaning tended to be stronger than those for the non-intended meaning, and this was true for both literal and non-literal usages.

34 Of the ten non-literal usages, eight have stronger non-literal chains than literal chains. The first exception is a text about a diver who is "getting his feet wet" in the diving profession (see Example [4] below). Here the idiom is clearly used tongue in cheek, and the cohesion with the literal meaning is probably intentional. In the second exceptional text, the strong cohesion with the literal reading is probably accidental. The text contains a non-literal usage of *playing with fire*, and the main topic of the text deals with bombs and rockets, which both annotators marked as being weakly related to *fire*.

35 Of the ten texts with literal usages, nine have stronger literal chains than non-literal chains. The single exception is a **bounce** text about car racing, in which the annotators found weak links between the non-literal meaning and words like *boring*, *slow*, and *speed*.

36 The results also confirm our second hypothesis: that non-literal usages generally exhibit lower degrees of cohesion with the text containing them than do literal usages. However, for most non-literal usages, the annotators marked some words in the context as being related to the non-literal meaning. Hence, even non-literal usages participate in cohesive relations with the context. At the same time, there

4. All of these were later classified as weak links in the GS.

5. Again, all were classified as weak in the GS.

tend to be fewer of these than for literal usages, and the relations tend to be weaker and more indirect (as indicated by the lower inter-annotator agreement seen for non-literal chains, Table 5).

37 All in all, the results of our annotation study confirm the hypothesis that literal and non-literal usages can be distinguished based on the cohesive relationships they enter into with their texts: strong literal chains indicate literal usages, and if the non-literal chain is stronger than the literal one, it is more likely that the expression is being used non-literally.

2.3. Mixed literal and non-literal use

38 Most of the time, determining whether a given target expression is being used literally or non-literally is a straightforward task. However, we encountered several interesting cases which seem to combine literal and non-literal uses. In these cases, it is more difficult to pull apart the interactions between the two cohesion chains. Here we discuss two examples.

2.3.1. Metaphorical "literal" expressions

39 The passage in [3] is taken from one of the **wall** texts. In this case, the phrase is used in its literal sense, but situated in a rich metaphorical context.

[3] That movie was entertaining in an off the wall way. "If Lucy Fell" **bounces off the wall** and drops to the floor like a pound of old fish.

40 The first sentence of this passage uses the idiom *off the wall*, which may be related to the target expression *bounce off the wall* but clearly has a distinct meaning. This is then echoed (via repetition of the last three words) in the second sentence, where the target expression occurs in a pseudo-literal usage. We call this "pseudo-literal" because it is meant to evoke the image of something wet and smelly hitting a wall and sliding down it. In this case, though, it is the movie which is (metaphorically) said to be sliding down the wall.

2.3.2. Signaling mixed use

41 Other interesting cases arise when the writer selects an idiom whose literal meaning relates to the topic of the text. One of the **feet** texts is about a man changing careers from drifter to diver. The target expression is used non-literally in reference to one of his early diving jobs. This is one of the exceptions to the general rule that the chain for the primary intended meaning should be stronger. The literal chain here contains 21 tokens, while the non-literal chain has only six.

[4] Davila said he worked for Disney to **get his feet wet**, so to speak.

42 In this case the phrase *so to speak* is used to draw attention to the nature of the use of the idiom, which has the flavor of a pun, due to its semantic proximity to the topic of the article. The phrase is one way of signaling a "complex" usage of

the idiom, where the main meaning is non-literal but there is also strong lexical cohesion between the text and the literal meaning of the expression.

43 In future work it would be interesting to explore the role and distribution of *so to speak* and similar cue phrases (e.g., *if you will*, *in a manner of speaking*, *as it were*). On cursory examination, such phrases often occur as a rhetorical strategy to express the author's awareness of the potential for multiple interpretations of the expression of interest, and perhaps also to call the reader's attention to that potential. One interesting question is whether they may also serve to point toward the topic of the text, suggesting that the literal meaning is a prominent theme in the text.

3. Automatic methods

44 Results from the human annotation study suggest that literal and non-literal usages of an expression can indeed be distinguished from one another on the basis of the strength of their cohesive links with the surrounding text. This result is potentially useful for automatic idiom detection, but only if such cohesive links can themselves be identified automatically. Because lexical cohesion is a matter of close semantic ties between words, automatic identification of cohesive links requires a measure of semantic relatedness that can be computed automatically for pairs of words.

3.1. Identifying cohesive links automatically

45 Modeling semantic relatedness is, of course, a very active research area in computational linguistics, and various relatedness measures have been proposed and used in previous research. We chose a measure called *NGD* (Cilibrasi & Vitanyi, 2007)[6], both because it has been previously used in an idiom detection task (Sporleder & Li, 2009) and because it has the advantage of not being restricted to classical relations. NGD computes relatedness on the basis of page counts returned by an internet search engine. The basic idea is that the more often two terms occur together relative to their overall frequency of occurrence, the more closely related they are. NGD is defined as follows:

$$[5] \qquad NGD(x, y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - min\{\log f(x), \log f(y)\}}$$

where x and y are the two words whose association strength is computed (e.g., *fire* and *coal*), $f(x)$ is the page count returned by the search engine for the term x

6. Cilibrasi and Vitanyi named their measure *NGD*, short for *Normalized Google Distance*, to reflect the fact that the measure takes into account page counts returned by a search engine. The name of this metric is essentially a generic use of "Google" as a stand-in for any major web search engine. There is no direct association between this research and the company of the same name. It is irrelevant which search engine is used. In our study, we did, in fact, use Yahoo! because we found its page counts to be more reliable and stable over time (see Sporleder & Li, 2009).

(and likewise for *f (y)* and *y*), *f (x, y)* is the page count returned when querying for *x AND y* (i.e., the number of pages that contain both *x* and *y*), and *M* is the number of web pages indexed by the search engine. Note that NGD is a measure of *distance*, where greater semantic similarity between two words results in a *lower* value for NGD.

46 It is important to note that the only purpose of making use of a search engine in this way is to collect statistics of word co-occurrence frequencies. This could also be done by using a large text corpus. However, even the largest corpus is significantly smaller than the amount of English language data available on the internet. Corpora typically have coverage problems: low frequency words may not occur often enough to compute reliable statistics and many words will be missing completely[7]. This applies especially, but not only, to proper names. For instance, it is impossible to compute the connection between the British football player Wayne Rooney and the word *ball* from the *British National Corpus* because *Wayne Rooney* does not occur in there. However, if a search engine is used instead to collect statistics from all English language text on the internet, the semantic relatedness between both expressions can be detected by the model. There is a strong tradition in computational linguistics of using search engine page counts as proxies for co-occurrence statistics collected from corpora and it has been shown empirically that statistics computed from page counts are at least as reliable and useful as those computed from text corpora (Lapata & Keller, 2005). However, nothing in our model hinges on the availability of a search engine; the model as such would also be usable with co-occurrence statistics computed by other means.

47 While search engine page counts have been found to generally produce accurate statistics, there are situations in which a search engine might produce erroneous counts. One known problem is that search engines do not always produce reliable page counts for high-frequency words (see Sporleder & Li, 2009). For this reason, we were not able to compute the cohesive structure for *play with fire*, as the search engine did not produce reliable numbers for either of the anchor words.

48 NGD was used in previous work (Sporleder & Li, 2009) to compute cohesion of the target expression with the literal chain only. A literal usage was predicted when cohesion was strong with the literal chain (i.e. low value for NGD), and non-literal usage was predicted otherwise, i.e., non-literal usage was treated as the default. In this study, we used NGD to measure cohesion with both the literal and the non-literal chain.

---

7. Of course, the entirety of the texts available on the internet is also not a "complete" representation of the English language. Many words will also be missing from this data source or occur very infrequently. There is also a certain amount of noise, e.g., ungrammatical sentences produced by non-native speakers. However, the internet is still significantly larger than any other available text corpus and the benefits to be gained from the sheer amount of data tend to outweigh the disadvantages associated with the presence of some noise.

Idiom	Paraphrases
bounce off the wall	"high-strung", "energetic", "over excited"
get one's feet wet	"first experience", "dabble", "dabbling"
rock the boat	"upset conventions", "break norms", "cause trouble", "disturb balance"
break the ice	"ease tensions", "get people talking", "facilitate communication"
play with fire	"risky behaviour", "risky behavior", "take risks", "act dangerously"

Table 6. Paraphrases for non-literal meanings

– **Cohesion with literal reading**. Because the meaning of a literally-used expression is essentially compositional, modeling cohesion of a literal occurrence of a potentially non-literal expression with the surrounding text is straightforward. We compute the NGD values between the content words of the target expression (e.g., *break* and *ice*) and all the other content words in the text.

– **Cohesion with non-literal reading**. The meaning of non-literal expressions is more difficult to model than that of literally-used expressions, because the component words do not serve as a reliable representation of the semantics of the expression. In this study, we compare two methods for computing cohesive links for non-literal meaning: (i) using the full string of the target expression, and (ii) using human-generated paraphrases of the non-literal meanings.

49 Under the first approach, we compute NGD between the full string (as *x*) and each content word of the text. The motivation for using the full string of the target expression is based on a study by Riehemann (2001), who found that expressions in canonical form (i.e., the dictionary form of an idiom) are more likely to be used non-literally than literally. Hence, while the pages returned by querying for the full string of the target expression (i.e., the canonical form) may contain some literal usages, the majority of pages should contain non-literal usages.

50 As expected, querying for the full string gets relatively low page counts since the frequency of the full expression is usually much lower than that of its parts. We also find that non-literal readings tend to appear in rather diverse contexts. For instance, *rock the boat* can mean "cause trouble" or "go against conventions". Words such as *accusation*, *attack*, and *conflict* are likely to co-occur with the first reading, while a different set of words, such as *counterculture*, *rebels*, *change*, *norm*, may co-occur with the second reading. The diversity of nuances in the non-literal meaning leads to a scattered distribution of the non-literal meaning across many different context words. As a result, the non-literal NGD is generally high (i.e., words tend to be rated as not very similar to the non-literal meaning). This actually closely resembles human intuition, in that humans also rate cohesive links with non-literal meanings as relatively weak.

51 Our second approach to modeling non-literal meaning uses human-generated paraphrases when querying the search engine. More precisely, we calculate NGD between the content words of the text and a set of paraphrases (see Table 6), using the logical operator **OR** to represent the full set of possible paraphrases (i.e., each query ranges over all strings). For better coverage, we intentionally use short expressions in the paraphrases. Intuitively, this method should lead to better results as paraphrases represent the semantic content of an idiom much more precisely than do either the component words or the full expression.

52 Comparing the results obtained by using the full-string model to those of the paraphrase model, we found evidence that the latter is more suited to modeling non-literal meaning. Using paraphrases generally leads to lower NGD values, i.e., more words from the text are rated as being semantically related to the non-literal meaning. Furthermore, the words rated as similar to the target meaning seemed more plausible than those returned by the full-string model. We thus used the paraphrase model in our final experiments described below.

3.2. Manually vs. automatically identified cohesive links

53 In our final experiment, we compared the cohesive links in the manually created GS to those found automatically by the method described above. Figures 2 to 5 plot the NGD for a given word against its position in the text. This allows us to see whether there are more and stronger cohesive links with words in the local vicinity of the target expression. The position of the target expression in the text is marked by a (blue) vertical line. Words that were marked as semantically related in the GS are indicated by a (green) bullet. Figure 2 shows the results for the literal chain of a literal usage of *rock the boat*, while Figure 3 shows the results for the non-literal chain for the same literal usage of *rock the boat*. Similarly, Figures 4 and 5 show the chains for a non-literal usage of *rock the boat*; the former depicts the non-literal chain, i.e., the chain for the intended usage, while the latter shows the literal chain.

Figure 2. Example of a literal chain for a literal usage ("rock the boat"). The *x* axis represents the position of the tokens in the text. The *y* axis is the NGD value between the token and the literal reading of the target expression (MWE)

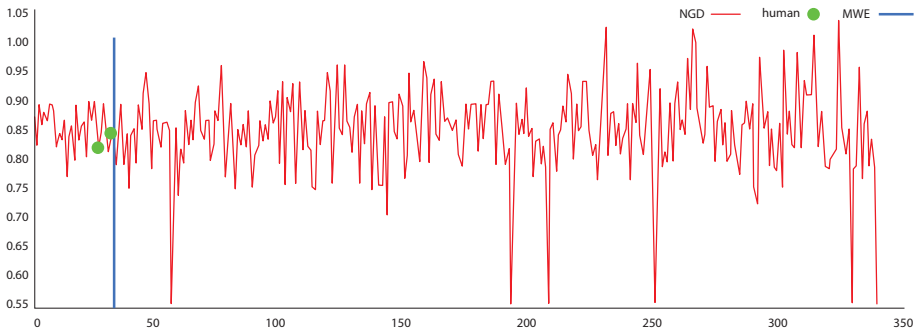

Figure 3. Example of a non-literal chain for the same literal usage as Figure 2 ("rock the boat"). The *x* axis represents the position of the tokens in the text. The *y* axis is the NGD value between the token and the non-literal reading of the target expression (MWE)

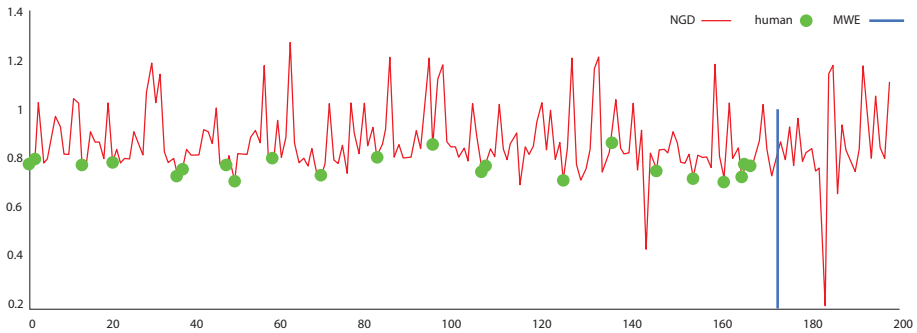

Figure 4. Example of a non-literal chain for a non-literal usage ("rock the boat"). The *x* axis represents the position of the tokens in the text. The *y* axis is the NGD value between the token and the non-literal reading of the target expression (MWE)

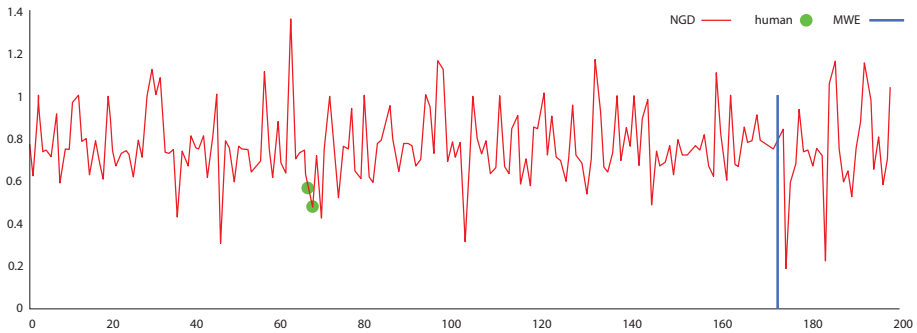

Figure 5. Example of a literal chain for the same non-literal usage as Figure 4 ("rock the boat"). The *x* axis represents the position of the tokens in the text. The *y* axis is the NGD value between the token and the literal reading of the target expression (MWE)

54 The first observation that can be made is that the position of a word in the text relative to the target expression does not seem to correlate with its likelihood to form a cohesive link, i.e., closely-related words tend to be scattered throughout the text and do not just appear in the neighborhood of the target expression. This is true both for the human annotation (i.e., there are several links with words far away from the target expression), and for the automatically computed NGD (i.e., the NGD is not necessarily lower in the vicinity of the target).

55 Second, it can be seen that human annotations agree quite well with the NGD values; words marked by humans tend to be located at local minima in the graph. Humans thus often mark those words whose NDG is relatively small, i.e., words which are rated as semantically similar to the target expression. This general pattern is observable for both the non-literal and the literal usages. Hence, it appears that computing NGD between the text and human-generated paraphrases is an effective strategy for modeling cohesion with non-literal expressions.

56 The results confirm that literal cohesion is stronger than non-literal cohesion. While most words in the literal chain have an NGD value below 0.5, most in the non-literal chain have an NGD value around 0.8. This is also in line with our findings for the human study.

57 The results also show that human judges tend to annotate more words in the literal chain than the non-literal chain (see Figure 2 vs. 3 and Figure 4 vs. 5). We also find that if there are only a few words in the non-literal chain, they are more likely to appear in the neighborhood of the target expression (see Figure 3). This suggests that a relatively small context may suffice for obtaining a measure of cohesiveness with non-literal readings.

4. Conclusions

58 In this study, we addressed the question of how non-literal and literal meanings participate in the cohesive structure of a text. Our findings suggest that both literal and non-literal meanings exhibit lexical cohesion with their context, however for non-literal meanings the cohesive ties tend to be much weaker. Links with the non-intended reading of an expression are typically weak, hence the cohesive structure of a text can be used to distinguish literal and non-literal readings. One exception arises in cases where an idiom is used tongue in cheek, i.e., it is deliberately chosen to cohere with both meanings.

59 We also investigated whether cohesive chains can be computed automatically. We found that a distance measure based on internet search engine page counts produces good results, i.e., it correlated well with human judgments. Furthermore, it seems that with this method the non-literal meaning of an expression can be modeled well by human-generated paraphrases.

60 In ongoing work, we are annotating a larger data set to explore the cohesive links in texts more fully. We are particularly interested in those cases where a deliberate play with words on the part of an author means that an expression exhibits cohesive links under both the literal and non-literal reading.

Acknowledgments

61 This work was funded by the German Research Foundation DFG within the Cluster of Excellence "Multimodal Computing and Interaction" (MMCI). Our thanks to the anonymous reviewers for very useful and insightful feedback; some points have been addressed here, and others remain for future work.

References

Baldwin, T. et al. 2004. Road-Testing the English Resource Grammar over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Stroudsburg: Association for Computational Linguistics: 2047-2050. Available online: http://www.lrec-conf.org/proceedings/lrec2004/pdf/754.pdf.

Barzilay, R. & Elhadad, M. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL-97 Intelligent Scalable Text Summarization Workshop (ISTS-1997)*. Stroudsburg: Association for Computational Linguistics: 10-17. Available online: http://www.aclweb.org/anthology/W/W97/W97-0703.pdf.

Beigman Klebanov, B. & Shamir, E. 2005. *Guidelines for Annotation of Concept Mention Patterns*. Technical Report 2005-8. Jerusalem: The Hebrew University of Jerusalem, Leibniz Center for Research in Computer Science.

Beigman Klebanov, B. & Shamir, E. 2006. Reader-Based Exploration of Lexical Cohesion. *Language Resources and Evaluation* 40 (2): 109-126.

Cilibrasi, R.L. & Vitanyi, P.M.B. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19 (3): 370-383.

Cramer, I. et al. 2008. Experiments on Lexical Chaining for German Corpora: Annotation, Extraction, and Application. *Journal for Language Technology and Computational Linguistics* 23 (2): 34-48.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hearst, M.A. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics* 23 (1): 33-64.

Hirst, G. & St-Onge, D. 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. Cambridge (Mass.): MIT Press: 305-332.

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hollingsworth, B. & Teufel, S. 2005. Human Annotation of Lexical Chains: Coverage and Agreement Measures. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (Beyond Bag of Words) – In Association with the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 26-32. Available online: http://research.yahoo.com/workshops/electra2005/electra-sigir-2005-proceedings.pdf.

Lapata, M. & Keller, F. 2005. Web-Based Models for Natural Language Processing. In *ACM Transactions of Speech and Language Processing* 2 (1): 1-31.

Morris, J. & Hirst, G. 2004. Non-Classical Lexical Semantic Relations. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*. Stroudsburg: Association for Computational Linguistics: 46-51. Available online: http://aclweb.org/anthology-new/W/W04/W04-2607.pdf.

Morris, J. & Hirst, G. 2006. The Subjectivity of Lexical Cohesion in Text. In J.G. Shanahan, Y. Qu & J. Wiebe (eds.), *Computing Attitude and Affect in Text*. Berlin: Springer.

Okumura, M. & Honda, T. 1994. Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In *COLING 1994: The 15th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. Vol. 2: 755-761. Available online: http://aclweb.org/anthology-new/C/C94/C94-2121.pdf.

Riehemann, S. 2001. *A Constructional Approach to Idioms and Word Formation*. PhD thesis. Stanford University, Department of Linguistics.

Sporleder, C. & Li, L. 2009. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*. Stroudsburg: Association for Computational Linguistics: 754-762. Available online: http://aclweb.org/anthology-new/E/E09/E09-1086.pdf.

Stührenberg, M. et al. 2007. Web-Based Annotation of Anaphoric Relations and Lexical Chains. In *Proceedings of the Linguistic Annotation Workshop*. Stroudsburg: Association for Computational Linguistics: 140-147. Available online: http://aclweb.org/anthology-new/W/W07/W07-1523.pdf.

Tanskanen, S.-K. 2006. *Collaborating towards Coherence: Lexical Cohesion in English Discourse*. Amsterdam: J. Benjamins.