*MAD 2010, Moissac, 17–20 march 2010*

# Cohesive Links with Literal and Idiomatic Expressions in Discourse: An Empirical and Computational Study

Caroline Sporleder, Linlin Li, and Alexis Palmer
Saarland University, Postfach 15 11 50
66041 Saarbrücken, Germany
{csporled, linlin, apalmer}@coli.uni-saarland.de

**Abstract.** Lexical cohesion is an important device for signalling text organisation. In this paper, we investigate to what extent idiomatic expressions participate in the cohesive structure of a text. We look at the problem both from an empirical and a computational perspective. Our results show that both literal and nonliteral expressions exhibit cohesion with their context, though the latter tend to do so to a lesser extent. We also show that cohesive links identified by humans can be approximated by an automatically computable semantic relatedness measure based on search engine page counts.

**Keywords.** idioms, multi-word expressions, cohesion, semantic relatedness, lexical chains, natural language processing, annotation

## 1 Introduction

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. In this paper we focus on lexical cohesion, which covers various kinds of semantic relationships between the lexical items in a text. These range from literal repetition (called 'reiteration' by Halliday and Hasan) to weaker semantic relationships (so-called 'non-classical relations' (Morris and Hirst2004)), as between *wet* and *dry* or *laugh* and *joke*.

Lexical cohesion is important because it is typically the most frequent class of cohesive ties (Hoey1991). It is also interesting because it tends to be global, i.e., texts can be analysed in terms of *chains* of lexically cohesive words. These chains can 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 indicate the overall structure and organisation of a text, e.g., in terms of the main topics addressed and their distribution throughout the text. Lexical cohesion has received considerable attention from both the linguistics (Hoey1991; Tanskanen2006) and the computational linguistics communities. Computational applications that make use of lexical chains range from the detection of malapropisms (Hirst and St-Onge1998) over word sense disambiguation (Okomura and Honda1994) and topic segmentation (Hearst1997) to automatic text summarisation (Barzilay and Elhadad1997).

In this paper, we address a particular aspect of lexical cohesion, namely how idiomatic expressions fit into the cohesive structure of a text. We are interested in (i) whether it is possible to find

cohesive ties between a figuratively-used expression such as *break the ice* 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 modelled computationally. The latter point is important because recent work by Sporleder and Li (2009) suggests that the cohesive structure of a text can provide cues for the detection of non-literal language. We evaluate whether the cohesive links found automatically are identical or at least similar to those annotated by humans. We 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.

# 2 Human evaluation of cohesive chains

To identify cohesive links, we carried out a small-scale annotation study, using texts from Sporleder and Li's (2009) dataset. We chose five expressions which can be used literally as well as idiomatically: *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 expressions were chosen for different reasons: for *bounce off the wall*, Sporleder and Li's method erroneously classified many literally used examples as non-literal, and the reverse is true for *get one's feet wet*. *Rock the boat* was included because the performance of the classifier was relatively high for this expression; *break the ice* and *play with fire* were selected more or less randomly. For each expression, we randomly chose four texts from the corpus for annotation: two with literal uses and two with non-literal uses. Two annotators labeled the complete texts, and a third annotator labeled the portion of text immediately surrounding the expression of interest, with a window of approximately two paragraphs in each direction.

To test the hypothesis that literal and non-literal meanings of an expression can be distinguished on the basis of lexical chains, we annotated two chains for each text: one for the literal meaning of the target expression (**literal chain**) and one for the idiomatic meaning (**idiomatic chain**). 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 authors try to deliberately minimise it to avoid confusion.[1] Our second hypothesis is that idiomatic usages tend to exhibit weaker cohesion (with the idiomatic chain) than literal usages do (with the literal chain).

## 2.1 Annotation process and decisions

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 generally report a relatively low inter-annotator agreement (Morris and Hirst2005; Hollingsworth and Teufel2005; Klebanov and Shamir2006). To alleviate this problem to an extent, the annotators in our study discussed some of the early texts post-annotation, with the

[1] Sometimes one can observe intended cohesion with both meanings, usually due to a deliberate play with words.

| 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 1: **Literal chains:** Number of cohesive links to anchor words of target expressions

aim of detecting potential problems and arriving at some general guidelines for the task. For this initial study, we include those texts in the results discussed below.[2]

After some discussion, the decision was made to mark 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, and named entities were also left unmarked. In addition, only content words were considered eligible for participation in cohesive chains.

**Literal chains.** Annotators identified **literal** and **idiomatic** chains for all texts, regardless of whether the target expression itself is used with its literal or its idiomatic 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 anchor words and the number of links to each appear in Table 1. 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.

**Idiomatic chains.** For the idiomatic chains, annotators marked words exhibiting lexical cohesion with the idiomatic 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 3).

**Gold standard.** A gold standard set of annotations was produced by adjudication over the individual annotations. We distinguished 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*.

**Multi-word expressions** The human annotators marked relevant multi-word expressions (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 will be ignored in the automated processing. (1) is from

[2]Table 2 shows the respective annotator agreement levels for discussed and undiscussed texts.

| Texts | Literal chains | Nonliteral chains |
|---|---|---|
| ALL | 0.8115 | 0.7354 |
| Literal usage | 0.8189 | 0.6724 |
| Nonliteral usage | 0.8031 | 0.7747 |
| Lit.discussed | 0.8142 | 0.6639 |
| Lit.not.disc | 0.8235 | 0.7061 |
| Nonlit.discussed | 0.9594 | 0.9388 |
| Nonlit.not.disc | 0.6859 | 0.6763 |

Table 2: Correlations between annotators 1&2

a text with an idiomatic 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.'

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 idiomatic meaning.

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**.

## 2.2 Findings

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 R. The top half of Table 2 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.

It can be seen that the correlation is generally relatively 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 is easier to agree on related words for literal usages, while non-literal usages are fuzzier and therefore less easy to annotate.

Once the human annotations had been adjudicated and a gold standard 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.[3] As expected, the chains for the intended meaning tend

[3]It should be noted that this measure does not make any distinction between the strength of individual links. In other words, weak and strong links count equally when determining strength of the chain.

to be stronger than those for the non-intended meaning, and this is true for both literal and non-literal usages.

Of the ten idiomatic 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*.

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*.

The results also confirm our second hypothesis: that idiomatic usages generally tend to exhibit lower degrees of cohesion with the text containing them than do literal usages. However, for most idiomatic usages, the annotators marked some words in the context as being related to the non-literal meaning. Hence, even idiomatic usages participate in cohesive relations with the context. At the same time, there 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 2).

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 idiomatic chain is stronger than the literal one, it is more likely that the expression is being used idiomatically.

## 2.3 Mixed literal and non-literal use

Most of the time, determining whether one of the target expressions is being used literally or idiomatically 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.

**Metaphorical 'literal' expressions.** 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.

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.

**Signalling mixed use.** 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 idiomatically 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.

In this case the phrase *so to speak* is used to draw attention to the nature of the use of the idiom, which almost has the feeling of a pun, due to its semantic proximity to the topic of the article. The phrase is one way of signalling 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.

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

Our results from the human annotation study suggest that literal and idiomatic usages of an expression can indeed be distinguished on the basis of their cohesive links with the surrounding text. However, this is only useful for automatic idiom detection if such cohesive links can be identified automatically. To be able to do so requires a measure of semantic relatedness that can be computed automatically for pairs of words. Modelling semantic relatedness is 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 *Normalized Google Distance* (NGD, see (Cilibrasi and Vitanyi2007)), since it has been used in an idiom detection task before (Sporleder and Li2009) and 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 occurrence frequency the more closely related they are. NGD is defined as follows:

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

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 $x$ (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*, i.e., a low value means that two words are rated to be very similar.

**Cohesion with literal reading.** To model the text's lexical cohesion with the literal usage, we computed the NGD between the content words of each target expression and all other content words in the text. A known drawback of measures like NGD is that search engines do not always produce reliable page counts for high-frequency words (see Sporleder and Li (2009)). For this reason, we were not able to compute the cohesive structure for *play with fire*, as the search engine had trouble with both anchor words.

**Cohesion with nonliteral reading.** Sporleder and Li (2009) computed cohesion with only the literal chain and then predicted literal usage if this chain was strong and idiomatic usage otherwise. For this study, we are interested in also computing cohesion with the nonliteral chain.

The meaning of a literally used expression is compositional and thus relatively easy to model. The idiomatic meaning is more difficult to model since an explicit semantic representation is missing. In our experiment, we compared two methods for computing cohesive links for nonliteral meaning: (i) by using the full string of the target expression, and (ii) by using human-generated paraphrases of the idiomatic meanings.

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 idiomatically 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 idiomatic usages.

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 found that idiomatic readings tend to appear in rather diverse contexts. For instance, *rock the boat* can mean *cause trouble* or *go against conventions*. It is more likely that words such as *accusation, attack, conflict* co-occur with the first reading, while words such as *counterculture, rebels, change, norm* co-occur with second reading. The diversity of nuances to the idiomatic meaning leads to a scattered distribution of the idiomatic meaning across many different context words. As a result, the nonliteral NGD is generally high (i.e., words tend to be rated as not very similar to the idiomatic meaning). This actually closely resembles human intuition, in that humans also rate cohesive links with idiomatic meanings as relatively weak.

In addition to using the full string to model idiomatic meaning, we also employed human-generated paraphrases, which were then used instead of the full string when querying the search engine. Intuitively, this method should lead to better results as paraphrases make it possible to pinpoint the meaning of an idiom more precisely. Table 3 shows the paraphrases we used. We deliberately tried to use short expressions in the paraphrases. Computationally, we represent the idiomatic reading by using the **OR** logic operator to connect all the possible paraphrases when sending a query to the search engine.

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 modelling idiomatic meaning. Using paraphrases generally leads to lower NGD values, i.e., more words from the text are rated as being semantically related to the idiomatic 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 experiment described below.

| 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 3: Paraphrases for idiomatic meanings

## 3.1 Manually vs. Automatically Found Cohesive Links

In our final experiment, we compared the cohesive links in the manually created gold standard to those found automatically by the method described above. Figures 1 to 4 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 gold standard are indicated by a (green) bullet. Figure 1 show the results for the literal chain of a literal usage of *rock the boat*, while Figure 2 shows the results for the nonliteral chain for the same literal usage of *rock the boat*. Similarly, Figures 3 and 4 show the chains for an idiomatic usage of *rock the boat*; the former depicts the nonliteral chain, i.e., the chain for the intended usage, the latter shows the literal chain.

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., the related words tend to be scattered all over the text and do not just appear in the neighbourhood 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).

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., which are rated as semantically similar to the target expression. This general pattern is observable for both the idiomatic and the literal meaning. Hence, it seems that modelling idiomatic meaning by combining NGD with human-generated paraphrases is a good strategy.

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

The results also show that human judges tend to annotate more words in the literal chain than the nonliteral chain (see Figure 1 vs. 2 and Figure 3 vs. 4). We also find that if there are only a few words in the nonliteral chain, they are more likely to appear in the neighbourhood of the target expression (see Figure 4). This means that the nonliteral reading can be obtained within a relatively small context.

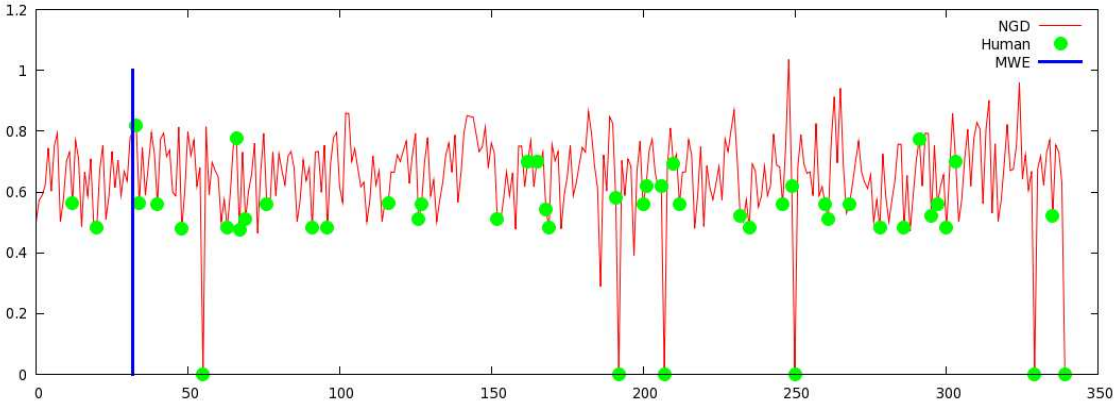


Figure 1: Example of a literal chain for a literal usage (“rock the boat”). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the literal reading of the target expression (MWE)

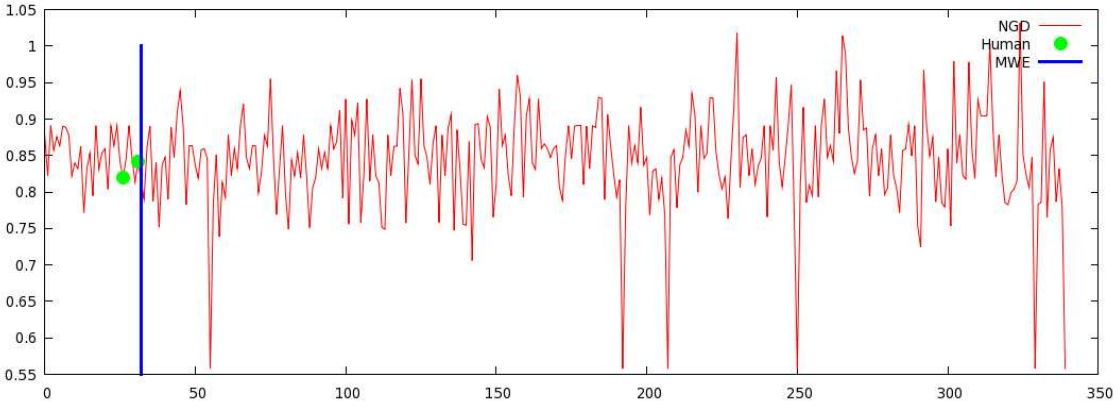


Figure 2: Example of a nonliteral chain for the same literal usage as Figure 1 (“rock the boat”). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the nonliteral reading of the target expression (MWE)

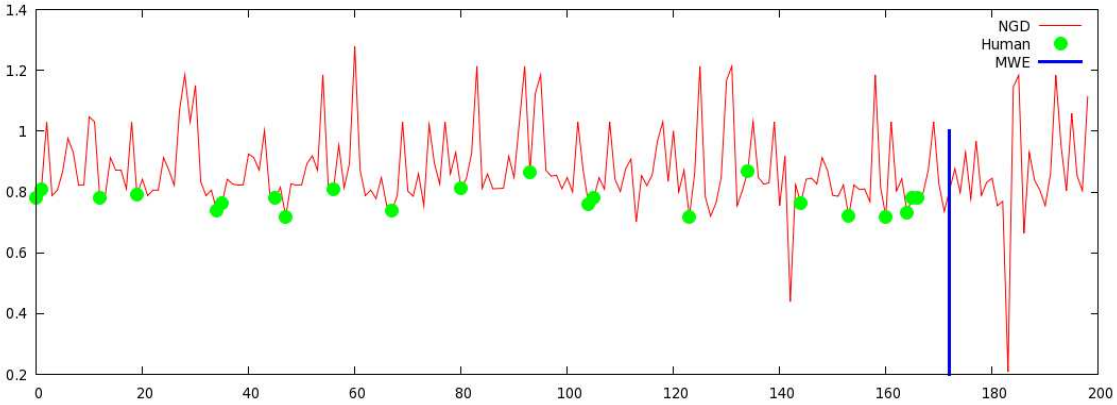


Figure 3: Example of a nonliteral chain for a nonliteral usage (“rock the boat”). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the idiomatic reading of the target expression (MWE)

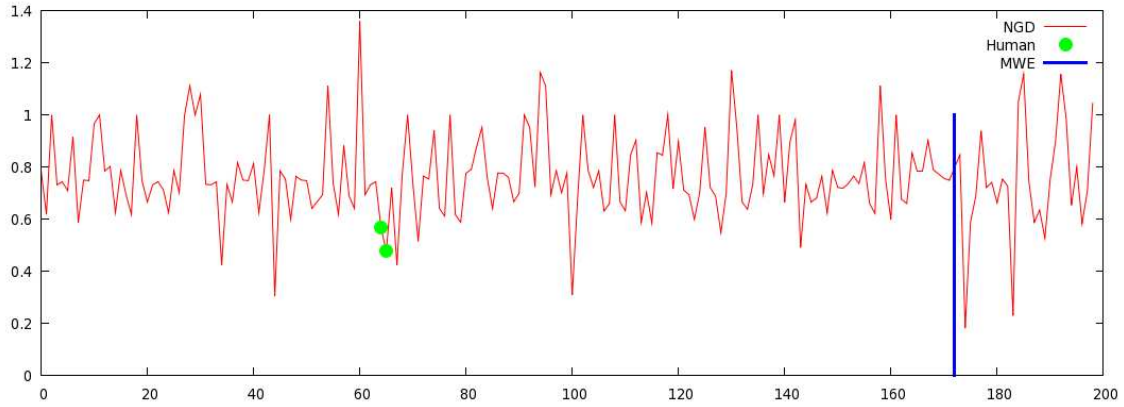


Figure 4: Example of a literal chain for the same nonliteral usage as Figure 3 ("rock the boat"). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the literal reading of the target expression (MWE)

# 4 Conclusions

In this study, we addressed the question of how idiomatic 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.

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. Furthermore, it seems that with this method the non-literal meaning of an expression can be modelled well by human-generated paraphrases.

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

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

# References

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for summarization. In *Proceedings of the ACL-97 Intelligent Scalable Text Summarization Workshop (ISTS'97)*.

Rudi L. Cilibrasi and Paul M.B. Vitanyi. 2007. The Google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19(3):370–383.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Marti Hearst. 1997. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. The MIT Press.

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

Bill Hollingsworth and Simone Teufel. 2005. Human annotation of lexical chains: Coverage and agreement measures. In *Proceedings of the SIGIR-05 Workshop ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications*.

Beate Beigman Klebanov and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40(2):109–126.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL-04 Workshop on Computational Lexical Semantics*, pages 46–51.

Jane Morris and Graeme Hirst. 2005. The subjectivity of lexical cohesion in text. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing attitude and affect in text*. Springer.

Manabu Okomura and Takeo Honda. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling-94)*.

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 the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*.

Sanna-kaisa Tanskanen. 2006. *Collaborating Towards Coherence: Lexical Cohesion in English Discourse*. John Benjamins.