Reader-based exploration of lexical cohesion

Beata Beigman Klebanov · Eli Shamir

Received: 22 June 2005 / Accepted: 18 August 2006 / Published online: 9 February 2007
© Springer Science+Business Media B.V. 2007

Erratum to: Lang Res Eval
DOI 10.1007/s10579-006-9004-6

Abstract Lexical cohesion refers to the reader-perceived unity of text achieved by the author's usage of words with related meanings (Halliday and Hasan, 1976). This article reports on an experiment with 22 readers aimed at finding lexical cohesive patterns in 10 texts. Although there was much diversity in peoples' answers, we identified a common core of the phenomenon, using statistical analysis of agreement patterns and a validation experiment. The core data may now be used as a minimal test set for models of lexical cohesion; we present an example suggesting that models based on mutually exclusive lexical chains will not suffice. In addition, we believe that procedures for revealing and analyzing sub-group patterns of agreement described here may be applied to data collected in other studies of comparable size.

Keywords Lexical cohesion · Inter-annotator agreement · Cohesion

1 Introduction

The quest for finding what it is that makes an ordered list of linguistic forms into a text that is fluently readable by people dates back at least to Halliday and Hasan's (1976) seminal work on textual cohesion, and gains in importance in applied

---

The online version of the original article can be found at <http://dx.doi.org/10.1007/s10579-006-9004-6>

B. Beigman Klebanov (✉) · E. Shamir
School of Computer Science and Engineering, The Hebrew University, 91904 Jerusalem, Israel
e-mail: beata@cs.huji.ac.il

E. Shamir
e-mail: shamir@cs.huji.ac.il

language processing community due to the advent of tasks like text generation¹ and text summarization.²

A number of “texturizing” elements were identified, studied and applied, including lexical repetition (Hearst, 1997; Hoey, 1991), patterns of entity realization (Barzilay & Lapata, 2005; Grosz, Joshi, & Weinstein 1995; Karamanis, Poesio, Mellish, & Oberlander, 2004; Siddharthan & Copestake, 2004), rhetorical organization (Mann & Thomson, 1988; Marcu, 2000).

The existence of **lexical cohesion**—texture created by using words with related meanings—has also been postulated (Halliday & Hasan, 1976). However, in contrast with such phenomena as lexical repetition, repeated reference using pronouns, or rhetorical structure often signalled by specific cue phrases, there are no clear form-based clues as to where lexical cohesion is found in the text. That is, no particular textual items are inherently lexically cohesive³; they are cohesive with respect to certain other lexical elements. Various classes of relations were suggested as carriers of cohesion (e.g. synonymy, hyponymy), but researchers noticed that much of lexical cohesive load is carried by relations *felt* by the reader, but difficult to classify (Halliday & Hasan, 1976; Hasan, 1984; Morris & Hirst, 2004).

The difficulty of identification and characterization of lexical cohesive relations places the reader in the center of the researcher’s attention. The fundamental question is how well readers agree on which items in a text are lexically cohesive. If the agreement is very poor, then the notion of lexical cohesion lacks internal coherence, and translates into something idiosyncratic for every individual reader. If there is substantial agreement, then we can try to develop identification procedures and characterization of those instances of lexical cohesion that are well agreed upon. These are likely to be a part of what the text is expected to induce upon an “average” reader, akin to the way repeated reference is systematically induced⁴, and thus a part of the textual structure.

Experimental, reader-based investigation of lexical cohesion is an emerging enterprise (Morris & Hirst, 2005). The main challenge is providing a translation of the notion of *patterns of lexical cohesion* into a task for the readers, such that (1) the results reflect researcher’s intuition about what he/she is after; (2) the guidelines are precise enough to enable replication but open enough to compel readers to use their intuition, which is the main source of information in an elicitation experiment.

Morris and Hirst asked people to group related words, to mark related pairs within each group, to name the relation, and to describe the meaning of each group. The researchers observed the average agreement of 63% on grouping, and lower agreement on identifying pairs of related words within each group.

We suggest that groups of words might emerge as a result of combining small pieces, rather than being primary units of analysis. Word groups are global structures; people might be tempted to make global decisions about the main issues in the

¹ See proceedings of SIGGEN meetings at <http://acl.ldc.upenn.edu/W/SIGGEN.html>.

² Workshops related to text summarization were held during annual meetings of the Association for Computational Linguistics every year since 2000; proceedings are available from <http://www.acl-web.org/anthology>.

³ Halliday and Hasan (1976) suggested that function words are not lexically cohesive; we address this issue in section 5.2.

⁴ See Hirschman, Robinson, Burger, and Vilain (1998), Poesio and Viera (1998) for evaluations of inter-reader agreement on co-reference structure.

text, and mark groups in light of those, not necessarily in a way sensitive to the cohesive impact of specific words in their particular place in the text.

For example, in Morris and Hirst's (2005) experimental text about movie characters and actors as possibly inappropriate role models for children, 4 people included *shooting* in the same group as *drinking*, which was entitled "bad behaviors"; one person, however, put *shooting* together with *police*, reflecting "law/order/authority" orientation.⁵ The question is whether the first placement of *shooting* was motivated by a global analysis of the text, where bad behaviors was perceived as a more salient issue than law and authority, whereas had people been given the freedom of marking pairwise relations without a need to form coherent groups, they would have connected *shooting* to both, or maybe just to the latter, as this could have made a stronger, readily perceivable connection, depending on the exact placement and rhetorical arrangement of the mentions of the three concepts.

In the following section, we present our version of a question to which lexical cohesion is an answer. Section 3 reports on the experiment we performed on 22 readers using this question. Section 4 contains analysis of inter-subject agreement and of the way it was used to identify a highly valid core of the phenomenon. Section 5 exemplifies and discusses the resulting structures. We conclude with an indication of potential usefulness of the resource created as a result of the experiment.

2 From lexical cohesion to anchoring

Lexical cohesive ties between items in a text draw on word meanings. Sometimes the relation between the members of a tie is easy to identify, like near-synonymy (*disease/illness*), complementarity (*boy/girl*), whole-to-part (*box/lid*), but the bulk of lexical cohesive texture is created by relations that are difficult to classify (Morris & Hirst, 2004). Halliday and Hasan (1976) exemplify those with pairs like *dig/garden*, *ill/doctor*, *laugh/joke*, which are reminiscent of the idea of scripts⁶ (Schank & Abelson, 1977): Certain things are expected in certain situations, the paradigm example being menu, tables, waiters and food in a restaurant.

However, texts sometimes start with descriptions of situations where many possible scripts could apply. A famous literary text starts with *Mother died today*. What are the generated expectations? A description of an accident, or of a long illness? A story about what happened to the family afterwards? The speaker's feeling of loneliness? Funeral arrangements? The mother's last wish and its fulfillment? Many directions are easily thinkable at this point.

We suggest that rather than generating predictions, scripts could provide a basis for abduction. Once any *normal* direction is actually taken up by the following text, there is a connection back to whatever makes this a normal direction, according to the reader's commonsense knowledge (possibly coached in terms of scripts). Thus, had the text developed the illness line, one would have known that it can be best

⁵ It is not clear whether the person who put *shooting* and *police* together included *drinking* in the same group as well.

⁶ Similar notions are *frame* and *schema*; see Bednarek (2005) for an overview.

explained-by/blamed-upon/abducted-to the previously mentioned lethal outcome. We will say in this case that *illness* is **anchored**⁷ by *died*, and mark it *illness* → *died*.

The cited line opens Albert Camus' novel *The Stranger* (1962), that proceeds thus:

Mother died today. Or, maybe, yesterday; I can't be sure. The telegram from the Home says YOUR MOTHER PASSED AWAY FUNERAL TOMORROW ...

The mention of *yesterday* refocuses the first sentence such that now it is seen as describing something that happened today, so continuing with recent history makes sense (*yesterday* → *today*). Later, *telegram* is seen in the light of death being an event that requires immediate reaction from relatives, so an urgent announcement is in order (*telegram* → *died*). Both these developments could hardly have been predicted from the first sentence—it is probably part of The Stranger's strangeness that he chooses to recount them rather than other things after such an opening, but the text does not lose cohesion. It is these kinds of connections—what is anchored by what—that we want to elicit from readers.

3 Experimental design

We chose 10 texts for the experiment: 3 news articles, 4 items of journalistic writing, and 3 fiction pieces. All news texts and one fiction story were taken in full; others were cut at a meaningful break to stay within 1000 word limit.⁸ The texts were in English—original language for all but two literary texts.

Our subjects were 22 students at the Hebrew University of Jerusalem, Israel; 19 undergraduates and 3 graduates, all aged 21–29 years, studying various subjects—computer science, cognitive science, biology, history, linguistics, psychology. Three participants named English their mother tongue; the rest claimed very high proficiency.

All participants first read the guidelines (Beigman Klebanov & Shamir, 2005) that contained an extensive example annotation, as well as short paragraphs explaining various technical matters (how to mark multiple and complex anchors), and highlighting some conceptual issues. In particular, people were asked to make an effort to separate personal knowledge from what they think is common knowledge, and general relations from instantial ones that are specifically constructed in the text using co-reference or predication.⁹ Participants then performed a trial annotation on a short news story, after which meetings in small groups were held for them to bring up any questions.

The experiment then started. For each of the 10 text, each person was given the text to read, and a separate wordlist on which to write down annotations. The wordlist contained words from the text, in their appearance order, excluding

⁷ The term is inspired by reference literature, especially regarding *bridging*—reference to an entity that is related to a previously mentioned entity, where the textual realization of this previous entity is called an anchor (Vieria & Poesio, 2000). Lexical cohesion, however, is not the same as referential cohesion in the text; appendix A and section 5.4 address this issue.

⁸ The experimental texts ranged between 330 and 982 words in length.

⁹ The exposition was non-technical and used examples. Appendix A reproduces the guidelines regarding general versus instantial knowledge.

the	be
stranger	sure #
albert	telegram
camus	from #
mother	home
died	says
today	your #
or	passed
maybe	away
yesterday	funeral
i	tomorrow
can't	...

Fig. 1 Beginning of the wordlist for *The Stranger*. Sharp signs mark places of repetition to ease reference back to the text. Two sharps mark repetitions of *the* and the third one—second occurrence of *mother*

verbatim and inflectional repetitions.¹⁰ Wordlists numbered between 175 and 339 items. For example, the beginning of *The Stranger* cited above corresponds to the wordlist in Fig. 1.

People were instructed to read the text, and then go through the wordlist and ask themselves, for every item on the list, which previously mentioned items help the easy accommodation of this concept into the evolving story, if indeed it is easily accommodated, based on the commonsense knowledge as it is perceived by the annotator. They were encouraged to use a dictionary if needed.¹¹ Figure 2 shows a possible annotation of the cited extract.

4 Analysis of experimental data

We now turn to a detailed presentation of data analysis; section 4.4 provides its summary.

Most of the existing research in computational linguistics that uses human annotators is within the framework of classification, where an annotator decides, for every test item, on an appropriate tag out of the pre-specified set of tags (Marcus, Santorini, & Marcinkiewicz, 1993; Poesio & Viera, 1998; Webber & Byron, 2004).

¹⁰ The exclusion was done mainly to keep the lists to a reasonable length while including as many newly mentioned items as possible. We conjectured that repetitions are usually anchored by the previous mention; this assumption is a simplification, since sometimes the same form is used in a somewhat different sense and may get anchored separately from the previous use of this form. This issue needs further experimental investigation.

¹¹ In our preliminary experiments with volunteers, people pointed out words they stumbled upon, like *arrondissement*. When preparing the texts for the experiment, we made a footnote near such items, and listed dictionary definitions of all their senses, in order to short-circuit likely dictionary lookups, without, however, taking from the reader the decision as to which sense(s) are relevant in the current context.

the	be
stranger	sure # $\rightarrow$ maybe
albert	telegram $\rightarrow$ died
camus $\rightarrow$ { albert stranger }	from #
mother	home $\rightarrow$ mother
died	says
today	your #
or	passed
maybe	away
yesterday $\rightarrow$ today	funeral $\rightarrow$ { passed_away died }
i	tomorrow $\rightarrow$ { yesterday today }
can't	...

Fig. 2 Possible annotation of the beginning of *The Stranger*. The notation $x \rightarrow \{c\ d\}$ means each of c and d is an anchor for x , and $x \rightarrow \{c\_d\}$ means c and d together anchor x

Although our task is not that of classification, we start from a classification sub-task, and use agreement figures to guide subsequent analysis. We use the by now standard κ statistic¹² (Carletta, 1996; Craggs & McGeeWood, 2005; Di Eugenio & Glass, 2004; Siegel & Castellan, 1998) to quantify the degree of above-chance agreement between multiple annotators, and α statistic for analysis of sources of unreliability (Krippendorff, 1980). The relevant formulas are given in Appendix B.

4.1 Classification sub-task

Classifying items into anchored/unanchored can be viewed as a sub-task in our experiment: Before writing any particular item as an anchor, the annotator asked himself whether the concept at hand is easy to accommodate at all. Agreement on this task averages $\kappa = 0.45$ (texts range from $\kappa = 0.36$ to $\kappa = 0.51$). These figures do not reach $\kappa = 0.67$, the accepted threshold for deciding that annotators were working under similar enough internalized theories¹³ of the phenomenon; however, the figures are high enough to suggest considerable overlaps.

Seeking more detailed insight into the degree of similarity of these theories, we follow the procedure described by Krippendorff (1980) to find outliers. We calculate the category-by-category co-markup matrix for all annotators¹⁴; then for all annotators except one, and by subtraction find the portion that is due to this one annotator. We then regard the data as two-annotator data (one versus everybody else), and calculate agreement coefficients. We rank annotators according to the

¹² Researchers working on discourse and dialogue found that the κ statistic is not always suitable, as the tasks are often not ‘pure’ classification tasks (see Artstein and Poesio (2005) for an overview). For example, (1) the categories are not equi-distanced—conflicting markups C_1 and C_2 could be less of a disagreement than C_1 and C_3 ; (2) there are cases of almost-agreement, where subjects need to locate something in the text and give very close answers, up to slight shifts in boundaries. Alternative methods of agreement measurements and/or adjustments of the κ -like statistics to such cases are actively researched (Artstein & Poesio, 2005; Passonneau, 2004). However, the anchored/unanchored subtask discussed in section 4.1 is a standard binary classification task, so the statistics apply.

¹³ whatever annotators think the phenomenon is after having read the guidelines

¹⁴ See formula 6 in Appendix B.

Table 1 Groups of annotators, according to conformity ranks

Group	Number of people	Average rank range	Agreement inside group (κ)
A	2	1.7–1.9	0.55
B	9	5.8–10.4	0.41
C	11	13.6–18.3	0.54

degree of agreement with the rest, separately for each text, and average over the texts to obtain the **conformity rank** of an annotator, between 1 and 22. The lower the rank, the less compliant the annotator.

Annotators' conformity ranks cluster into 3 groups shown in Table 1. Group A are consistent outliers—their average rank for the 10 texts is below 2. Group B are, on average, in the bottom half of the annotators with respect to agreement with the common, whereas members of group C display relatively high conformity.

It is possible that groups A, B and C have different interpretations of the guidelines, but our idea of the common (and thus the conformity ranks) is dominated by the largest group, C. Within-group agreement rates are shown in the last column of Table 1. The two annotators in group A seem to have an alternative understanding of the task, being much better correlated between each other than with the rest of the people; appendix C gives more details about their annotations.

The numbers for the other two groups could support two scenarios schematically depicted in Fig. 3: (a) each group settled on a different idea of the phenomenon, where group C is in better agreement on its version than group B on its own; (b) people in groups B and C have basically the same interpretation, but members of C are more systematic in carrying their idea through. It is crucial for our analysis to tell those apart—in the case of multiple stable interpretations it is difficult to talk about *the* anchoring phenomenon; in the core-periphery case, there is hope to identify the common core emerging from 20 out of 22 annotations.

If the two groups have different interpretations, adding a person p from group C to group B would usually not improve the agreement in the target group (B), since p is likely to have a different interpretation than B's members. If, however, the two groups have the same interpretation, moving p from C to B would usually improve the agreement in B, since, coming from a more consistent group, p 's agreement with the common interpretation is expected to be better than that of an average member of group B.

We performed this analysis on groups A and C with respect to B. Adding members of A to B improved the agreement in B only for 1 out of the 10 texts. Thus, the relationship between the two groups seems to be that of different interpretations. Adding members of C to B resulted in improvement in agreement in at least 7 out of 10 texts for every added member. Thus, the difference between groups B and C is that of consistency, not of interpretation; we may now search for the well-agreed-upon core of this interpretation. We exclude the two outliers (group A) from subsequent analysis. The remaining group of 20 annotators exhibits an average agreement of $\kappa = 0.48$ on anchored/unanchored classification, texts ranging from $\kappa = 0.40$ to $\kappa = 0.54$. The improvement in agreement after the exclusion of outliers is significant at $p < 0.01$ (Wilcoxon matched-pairs signed-ranks test, $n = 10$, H_1 : agreement scores improved after excluding the two outliers).

Fig. 3 Schematic patterns of agreement in two groups: circles versus squares. **(a)** Groups vary in interpretation and consistency; **(b)** Groups have the same interpretation, but vary in consistency

4.2 Finding the common core of the classification

We now seek a reliably classified subset of the data. The main concern is not including cases of agreement that could be due to chance with intolerably high probability. To estimate this probability, we induce 20 random *pseudo-annotators* from the 20 actual ones: Each pseudo-annotator marks the same proportion of items as anchored as the respective actual annotator, but chooses the items at random. We model this by letting the i -th pseudo-annotator toss a coin with $p(\text{heads}) = p_i$, independently for every item, where p_i is the proportion of items marked as anchored by the actual annotator i in the whole of the dataset; random variables x_i represent the outcomes of such tosses. A random variable S ranging between 0 and 20 says how many pseudo-annotators marked an item as anchored, having tossed each his own coin, independently of each other. The expectation and variance of S are given by $E(S) = \sum_{i=1}^{20} p_i$ and $V(S) = \sum_{i=1}^{20} p_i(1-p_i)$.

We assume that numerous repeated trials on S are normally distributed. We calculate the probability $p(\text{Val})$ that a single observation from a normal distribution with parameters $\mu = E(S)$, $\sigma = \sqrt{V(S)}$ will fall in the interval $(-\infty, \text{Val}]$. Calculating for $\text{Val} = 0.5$, we test how likely it is that none of the pseudo-annotators anchored a given item; $1 - p(12.5)$ gives the probability that at least 13 out of 20 pseudo-annotators anchored a given item.¹⁵

Now, $1 - p(\text{Val})$ is the confidence with which we can reject the hypothesis that certain observed agreement levels are due to chance, since S models chance agreement. We seek values that allow high confidence levels, both for anchoredness and unanchoredness decisions. Thus, with 99% overall confidence ($p < 0.01$), we may trust unanimous decisions on unanchoredness, and decisions of at least 13 out of 20 people that an item is anchored.¹⁶ Allowing 95% confidence on each of the two decisions separately, cases of at most 2 markups can be reliably considered as unanchored ($p(2.5) = 0.0256$), and anchoring by at least 11 people is enough for anchoredness ($1 - p(10.5) = 0.0290$). For the subsequent analysis, we choose the 99% confidence version.

Figure 4 plots the actual data for *The Stranger*, along with the reliability cutoffs.

4.3 Identifying anchors for anchored items

The next step is identifying reliable anchors for the anchored items. We calculated average anchor strength for every text: the number of people who wrote the same

¹⁵ Number of annotators is an integer, “nobody” means 0, but normal distribution has real-valued data. We use the standard technique of counting values $\leq x.5$ as x , and $> x.5$ as $x+1$.

¹⁶ $p(0.5) + (1 - p(12.5)) = 0.0018 + 0.0021 = 0.0039 < 0.01$

Fig. 4 Actual classification pattern for the beginning of *The Stranger*. The items *funeral*, *yesterday*, *tomorrow* are anchored with 99% confidence. The complex item *passed_away* was not in the original wordlist, but was added by 12 annotators as anchored

anchor for a given item, averaged on all reliably anchored items in a text. The average anchor strength is between 5 and 7 for the different texts. Taking only **strong anchors** (anchors of at least the average strength), we retain about 25% of all anchors assigned to anchored items in the reliably classified data. In total, there are 1261 pairs of reliably anchored items with their strong anchors, between 54 and 205 per text; we refer to this set as **core data**.

Strength cut-off is a heuristic procedure; some strong anchors were marked by as few as 6 out of 20 people, making it unclear whether they can be trusted as embodiments of the core of the anchoring phenomenon in the analyzed texts. We thus devised an anchor validation experiment, reasoning as follows: In the original experiment, people were asked to generate all anchors for every item they thought was anchored. In fact, people generated only 1.86 anchors per anchored item. This makes us think that people were most concerned with finding *an* anchor, making sure that something they think is easily accommodatable is given at least one preceding item to blame for that; they were less diligent in marking up all such items. We conjectured that judging presented anchors would be easier than finding ones, so in the validation experiment people were asked to cross over anchors they did not agree with.

17 of the 20 annotators participated in the validation experiment. We chose 6 out of the 10 texts; each person performed validation on 3 out of those, such that every text received 7–9 validation versions. For each text, readers were presented with the same list of words as in the first part, only now each word was accompanied by a list of anchors. For each item, every anchor generated by some reader was included; the order of the anchors had no correspondence with the number of people who generated it. A small number of items also received a random anchor—a randomly chosen word from the preceding part of the wordlist. Figure 5 shows such a list for the beginning of *The Stranger*.¹⁷

Ideally, if lack of markup is merely a difference in attention but not in judgment, all non-random anchors should be accepted. To see the distance of the actual results from this scenario, we calculate the total mass of votes as number of anchored-anchor pairs times number of people, check how many are accept votes, and average

¹⁷ *The Stranger* was not included in the validation experiment.

the	be → died
stranger	sure # → { can't maybe }
albert → the_stranger	telegram → died
camus → {albert the_stranger stranger}	from #
mother	home → {died mother}
died	says → telegram
today	your # → mother
or	passed → died
maybe → or	away → { passed died}
yesterday → {today maybe}	passed_away → died
i	funeral → {passed_away your_mother died}
can't	tomorrow → {yesterday today}

Fig. 5 Wordlist for validation experiment for the beginning of *The Stranger*. Core data annotations are *yesterday* → *today*, *funeral* → { *passed_away died*} and *tomorrow* → { *yesterday today*}

over texts. People accepted 62% of non-random pairs, 94% of core data pairs (texts scoring between 90% and 96%), and only 15% of pairs with a random anchor.

4.4 Summary of data analysis

We used an anchored/unanchored classification sub-task to establish the existence of common interpretation among 20 out of 22 annotators and to identify items at the reliably agreed-upon core of this interpretation.¹⁸ For items reliably classified as anchored, we identified strong anchors and validated those in an additional experiment: Even people who did not actually generate them, accepted them as valid anchors (94% average acceptance rate). We may thus regard the core data as relatively uncontroversial manifestation of anchoring patterns in the examined texts, reliably found by a group of 20 readers.

5 From anchoring to lexical cohesion

This section discusses the implication of the current study for modeling and annotation of patterns of lexical cohesion. First, we show that the pairwise links organize into interesting global patterns (section 5.1). We then discuss the part-of-speech composition of the cohesive pairs (section 5.2), the character of unreliable annotations (section 5.3), and the issue of multi-word items (section 5.4).

¹⁸ We experimented with inducing a direct classification task on anchored-anchor pairs: 1 if a person generated the pair, 0 if not. Average agreement was $\kappa = 0.34$; we chose not to use this task in subsequent analysis since we think it does not reflect agreement faithfully. Such classification regards sets of markups: $\{\{x \rightarrow y\}, \{x\}\}$, and $\{\{x \rightarrow \{yz\}\}, \{x \rightarrow z\}\}$ as agreeing on all pairs involving x apart from $x \rightarrow y$. However, in the first case, there is no agreement even on whether x is easy to accommodate at all, let alone on an anchor, whereas in the second case, there is agreement that x is anchored by z , but one annotator thinks it is also anchored by y , which we think is a much less severe disagreement.

5.1 Cohesion beyond word pairs

To exemplify the observed patterns, we organize the core data in a graph, where a downward arrow from *b* to *a* means that *a* is a strong anchor for *b*. Figure 6 shows the two largest connected components for the first 12 sentences of Jay Teitel's 1987 Toronto Magazine article titled *Outland* (shown as appendix D), reproduced in Morris and Hirst's (1991) and analyzed therein into **lexical chains** that "delineate portions of text that have a strong unity of meaning" (page 23). Numbers inside nodes correspond to Morris and Hirst's chain numbers; no number means a word was not assigned to any chain.

Inspecting the upper component, we see that its right-hand side is rooted in *driving* and the left-hand one in *afflicted*. Walking up the structure we notice that the connection between the two halves hangs on a single link, going from *lights* to *car*. Indeed, *lights* is anchored by *car*, by *blindness* and by *night*, which reflects the major rhetorical role played by *lights* in this text—that of connecting driving issues to environmental lack of light (darkness, dusk, night) and to human ailment (blindness, afflicted, deadly), as reflected in the following passage: "... I passed them [those years] driving ... in a Volkswagen afflicted with night blindness. The car's lights never worked ..."

In the second component we notice the pivotal position of *neighbourhood*, as a social entity (community, collective, people), as a kind of residential unit (city, suburbs, apartment), and as a physical place (environment, surroundings).

These examples undermine an assumption sometimes made in applied research that lexical cohesive patterns can be represented by mutually exclusive chains (Barzilay & Elhadad, 1997; Silber & McCoy, 2002; Stokes, Carthy, & Smeaton,

Fig. 6 Anchoring Patterns versus Lexical Chains

Table 2 Proportions of POS in wordlists and in the core anchoring data. *Other* category contains pronouns, conjunctions, prepositions, numbers, articles, etc

POS	Wordlist proportion		Core Data Proportion			
	Average	Range	As Anchored		As Anchor	
			Average	Range	Average	Range
Noun	0.32	0.27–0.34	0.67	0.54–0.79	0.64	0.46–0.77
Adj+V	0.30	0.24–0.39	0.29	0.10–0.44	0.30	0.16–0.51
Adv	0.10	0.04–0.14	0.02	0.00–0.06	0.02	0.00–0.06
PN	0.05	0.01–0.16	0.02	0.00–0.07	0.03	0.00–0.10
Other	0.23	0.19–0.28	<0.01	0.00–0.04	<0.01	0.00–0.03

2004), realized in Morris and Hirst's (1991) analysis as well.¹⁹ Putting every word in at most one chain misses important lexically expressed connections between meaning components that are registered by human readers.

5.2 Part-of-speech

Another issue in lexical cohesion research is the relationship between the part-of-speech (henceforth, **POS**) of an item and its participation in lexical cohesion: It is usually assumed that content words are solely responsible for the lexical cohesive texture (Halliday & Hasan, 1976; Morris & Hirst's 1991). Moreover, applied research tends to concentrate on the nominal texture (Al-Halimi and Kazman, 1998; Barzilay & Elhadad, 1997; Hirst & Budanitsky, 2005; Stokes et al., 2004). Table 2 shows the split of the wordlists and of the core data by POS.

The table indicates that nouns favor cohesive structures: they are twice more frequent within core items than in the wordlists. Adjectives and verbs keep to their proportion; adverbs and proper nouns tend not to participate in lexical cohesion. Function words were occasionally marked by the annotators, but not systematically enough to make it to the reliable data; core pairs with the dollar sign (\$) and with 747 (the number of a popular Boeing model) are rare exceptions.

The implication for an algorithm that would try to replicate core annotations is clear: by looking only at common nouns, verbs and adjectives, it is possible to ignore almost 40% of first mentions, and retain a chance for above 90% recall on the core data. For some texts, exclusion of proper names might cause a more significant recall loss, as in a news story where 10% of core anchors are proper names.

The implications for future annotation work are not so clear. We note the variability in proportions of the different POS both in the wordlists and in the core annotations: Whereas proper names are usually rather marginal, in one of the news texts they comprise 16% of wordlist items, and participate as anchors in 10% of core pairs in another news story. Similarly, adverbs usually carry little cohesion, but in some texts they are not so lightweight; in fact, *yesterday*, *today*, *tomorrow* form salient cohesive patterns in *The Stranger*.

As for function words, it is tempting, and might indeed be worthwhile, to cut the wordlists down in quarter, likely preserving (almost) all of the cohesive texture. A minor caveat has to do with the specific way in which data was presented in our

¹⁹ The only case to the contrary is *darkness* in both chains 2 and 8.

experiment: The annotators had the wordlist aligned with the actual text through the use of the sharp sign, as shown in Fig. 1. However, the more items are skipped, the more cumbersome and unhelpful such presentation will be, making the alignment with the text more difficult. The second caveat is of a more general nature: In an exhaustive annotation process, we might not want to miss even the little amount of texture that might be residing in the function words.

5.3 Unreliable annotations

The analysis in section 4.2 allows identification of the *least* reliable annotations as well: those are items marked as anchored by the number of people closest to the expectation of the variable S that models chance agreement, which is between 6 and 7 in our data. Such items, exemplified in Fig. 7, are neither reliably anchored nor reliably unanchored.

Checking these examples with respect to the text, we suggest that they might reflect an interplay between conflicting influences, such as textual distance, reference, syntax, and the non-specificity of a good anchor.

Thus, pairs like *vividly* → *remember*, *holes* → *punched* stand in syntactic dependency relations in the text, which could intensify their perceived cohesion. A similar influence of referential cohesion is likely in *ceiling* → *car*, *girls* → *names*, *girls* → *phone_numbers*, *suburbs* → *outland*: the text discusses the ceiling of a car, phone numbers and names of girls, and refers to *suburbs* as *outland* (this is a possible understanding of the title).

In other cases, it seems that the textual rendering of the concepts works against their cohesiveness—for example, *reminding* and *remember* are more than 150 wordlist items apart, so their cohesiveness, while still perceptible, is backgrounded.

Cases like *medieval*, *swore*, *certainty* are characterized by a large variety of proposed anchors, and by the small number of people who wrote each anchor. The annotators are likely to have vaguely felt that the item ought to be anchored in the text, but had a difficulty anchoring it to any specific preceding item.

We think that a possible interpretation of the unreliably marked cases is that lexical cohesion is not a binary phenomenon, and such cases are examples of an intermediate degree of cohesion: They are anchored in the text enough for some people to notice it, but the anchoring is not as salient as for items marked by a decisive majority of the annotators.

5.4 Multi-word items

Multi-word items are an interesting issue with respect to lexical cohesion. On the one hand, the split into atomic units, as done in the current experiment, seems rather

Fig. 7 Items from *Outland* marked as anchored by 6 or 7 people, with their anchors. The number of people who marked each anchor is shown in parentheses

vividly	remember(7)
holes	punched(6)
suburbs	outland(7) life(1)
ceiling	car(6) vinyl(2) holes(1) windows(1)
girls	phone_numbers(4) names(4) university(1)
reminding	remember(6) notice(2) get_used(1)
medieval	night(3) afflicted(2) years(2) blindness(1)
swore	penance(3) hating(2) wanting(1) never(1) deadly(1)
certainty	swore(2) relentless(2) surprising(1) get_used(1) relatively(1)

artificial in cases like *New York*; indeed, in applied research, phrases are sometimes included as units of analysis (Barzilay and Elhadad, 1997; Stokes et al., 2004).

On the other hand, the decision regarding the boundaries of complex items is not always easy to make. For example, is *New York City* a unit, or does *City* stand on its own? Is the *Journal* part of *Wall Street Journal* a separately functioning item, and is *Wall Street* such an item as well? In fact, the core data contains the pairs *business* → *wall_street* and *report* → *journal*, where the anchors come from *Wall Street Journal* in the relevant text.

The main reason for not including phrases in the wordlists is that phrases, as opposed to atomic units, are not marked orthographically in the text, but are in the readers' eyes. This means both that things often taken to be phrases could be perceived in separation as well (as in the *Wall Street Journal* example above), and that things that are compositional and thus are not good a-priori candidates for phrasehood, are sometimes perceived as a unit. For example, *last_minute* is a core-data anchor for *hurrying* in one of our texts.

In addition, we believe that permitting the split of phrases is useful in trying to pinpoint the part that carries the meaning component that is mostly responsible for the anchoring link. Barzilay and Elhadad (1997) default to the head noun of the compound as the part that is responsible for the cohesive behavior of the complex item; cases like the aforementioned *business* → *wall_street* suggest that this is not always justified.

Furthermore, the identity of a sequence of atomic items as a unit is often tied to its functioning as a single referring expression: It is possible that a text mentioning *Wall Street Journal* will keep discussing the journal, never mentioning *Wall Street* as a separate entity. There is thus a potential tension between referential and lexical structure—readers might or might not perceive parts of a single referring expression as separately functioning lexical units.

The handling of phrases is clearly an issue for future experimental research into lexical cohesion, and into its relationship with other structures in the text.

6 Conclusion

This article reports on a reader-based experiment that used common knowledge based conceptual anchoring to operationalize the notion of *lexical cohesion*. There was much diversity in peoples' responses. However, we developed statistics-based procedures, which, for this experiment, identified a common core of the phenomenon. This core was validated in an additional experiment with the same group.

The core data may now be used as a resource in computational exploration of lexical cohesion. It can serve as an intrinsic test set for models of lexical cohesion: Any good model should at least get the core part right. Currently, models of lexical cohesion are only evaluated in an extrinsic fashion—by their usefulness for a given application (Al-Halimi and Kazman, 1998; Barzilay & Elhadad, 1997; Green, 1998; Hirst & Budanitsky, 2005; Stairmand, 1997; Stokes et al., 2004). As we saw, models based on mutually exclusive lexical chains built along “meaning components” are expected to miss salient connections; such shortcomings can be detected using the intrinsic test, before any high-level application is attempted. To encourage such experimentation, the core data is publicly available at the first author's homepage.²⁰

²⁰ currently at <http://www.cs.huji.ac.il/~beata/>

An additional use of the data, suggested by an anonymous reviewer, is a sequel experiment in which subjects would be asked to classify the anchoring pairs according to the type of relation they embody. This would allow separation between the stage of identification of cohesive pairs and that of classification, which we believe is important, as merging the two might bias the annotators towards marking pairs with easy-to-classify relations. Such data would be useful for finding out which lexical relations participate in the cohesive structures, and to what extent.

Another contribution of this work is a demonstration of the use of agreement statistics not just for annotation reliability judgment, but also for a more detailed exploration of annotation patterns exhibited by a relatively large group of subjects. The development of resources related to semantic/pragmatic tasks is impeded by the difficulty to attain high inter-annotator agreement (Morris & Hirst, 2005; Poesio & Vieira, 1998). We believe that finer agreement analysis techniques will lead to more effective and insightful use of experimental data that exhibits robust agreement but not enough to warrant a conclusion that the whole of it has been reliably annotated.

Appendix

A Instantial vs. General Knowledge—Extract from Guidelines

A text is telling us a story about certain entities. For example, it could tell us that children went out in a boat with their father, there was a storm but they survived because the father was an experienced sailor. Whereas relating father back to children, storm to boat, survival to storm and sailor to boat would be based on the general knowledge of what a family is, where you go in a boat and what can happen there, a connection between sailor and father is not something general/typical, it is created only because in this particular text the two descriptions apply to the same person. That is, the relation is instantial, pertaining to a certain instance (a text), not a general one. Since we are after general knowledge, please do not mark instantial connections.

B Measures of Agreement

Let N be the number of items to be classified; m —the number of categories to classify into; k —the number of raters; n_{ij} is the number of annotators who assigned the i -th item to j -th category. We use Siegel and Castellan's (1998) version of κ ; it assumes similar distributions of categories across coders in that it uses the average to estimate the expected agreement (see equation 2). This characteristic has been both criticized (Di Eugenio and Glass 2004) and commended (Craggs and McGeeWood, 2005); in any case, the current experiment employs 22 coders, making averaging a much better justified enterprise than in studies with very few coders typical in discourse annotation work (Di Eugenio and Glass, 2004). The calculation of the α statistic follows Krippendorff (1980).

The K statistic

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

$$P(E) = \sum_{j=1}^m p_j^2, \quad p_j = \frac{\sum_{i=1}^N n_{ij}}{Nk} \quad (2)$$

$$P(A) = \frac{1}{Nk(k-1)} \sum_{i=1}^N \sum_{j=1}^m n_{ij}(n_{ij} - 1) \quad (3)$$

The α statistic

$$\alpha = 1 - \frac{D_{\text{obs}}}{D_{\text{exp}}} \quad (4)$$

$$D_{\text{obs}} = \frac{1}{\zeta} \sum_{j=1}^m \sum_{i=1}^m \beta_{ji}(1 - \Delta_{ji}), \quad \zeta = mk(k-1) \quad (5)$$

$$\beta_{ji} = \sum_{l=1}^N n_{lj}(n_{li} - \Delta_{ji}), \quad \Delta_{ji} = \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases} \quad (6)$$

$$D_{\text{exp}} = \frac{1}{\zeta(\zeta - k + 1)} \sum_{j=1}^m \sum_{i=1}^m \gamma_j \gamma_i (1 - \Delta_{ji}), \quad \gamma_i = \sum_{j=1}^m \beta_{ji} \quad (7)$$

C Outlying Annotations

The dissidence of the 2 members of group A could suggest that the guidelines left some room for a different understanding. These two annotators marked more than twice as many items as anchored as the average of the rest of the people (72% vs. 33%), including many instances that followed syntactic dependency relations. For example, a stretch of text saying “... I had an idea he looked annoyed, and I said, without thinking...” got analyzed by people in group A into the following relations: *idea* → *had*, *annoyed* → *looked*, *thinking* → *without*. They were the only ones to mark an anchor for *idea* at all; out of 12 other people who anchored *annoyed*, 11 named *refuse* as the anchor and just one had *looked*; 6 additional people who anchored *thinking* all named *idea* as its anchor. It seems that people from group A were overly influenced by syntax, sometimes at the expense of conceptual relations we were after.

D Extract from Outland, by Jay Teitel, 1987

I spent the first 19 years of my life in the suburbs, the initial 14 or so relatively contented, the last four or five wanting mainly to be elsewhere. The final two I remember vividly: I passed them driving to and from the University of Toronto in a red 1962 Volkswagen 1500 afflicted with night blindness. The car's lights never

worked—every dusk turned into a kind of medieval race against darkness, a panicky, mournful rush north, away from everything I knew was exciting, toward everything I knew was deadly. I remember looking through the windows at the commuters mired in traffic beside me and actively hating them for their passivity. I actually punched holes in the white vinyl ceiling of the Volks and then, by way of penance, wrote beside them the names and phone numbers of the girls I would call when I had my own apartment in the city. One thing I swore to myself: I would never live in the suburbs again.

My aversion was as much a matter of environment as it was traffic—one particular piece of the suburban setting: the “cruel sun”. Growing up in the suburbs you can get used to a surprising number of things—the relentless “residentialness” of your surroundings, the weird certainty you have that everything will stay vaguely new-looking and immune to historic soul no matter how many years pass. You don’t notice the eerie silence that descends each weekday when every sound it drained out of your neighbourhood along with all the people who’ve gone to work. I got used to pizza, and cars, and the fact that the cultural hub of my community was the collective TV set. But once a week I would step outside as dusk was about to fall and be absolutely bowled over by the setting sun, slanting huge and cold across the untreed front lawns, reminding me not just how barren and sterile, but how undefended life could be. As much as I hated the suburban drive to school, I wanted to get away from the cruel suburban sun.²¹

Reference

- Al-Halimi, R., & Kazman R. (1998). Temporal indexing through lexical chaining. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 333–351). Cambridge, MA: MIT Press.
- Artstein, R., & Poesio M. (2005). $\text{Kappa}^3 = \text{Alpha}$ (or Beta). Technical report, University of Essex CS Technical Report CSM-437.
- Barzilay, R., & Elhadad M. (1997) Using lexical chains for text summarization. In *Proceedings of the ACL intelligent scalable text summarization workshop* (pp. 86–90). Madrid, Spain.
- Barzilay, R., & Lapata M. (2005) Modeling local coherence: An entity-based approach. In *Proceedings of ACL-05*. Ann Arbor, USA.
- Bednarek, M. A. (2005). Frames revisited – The coherence-inducing function of frames. *Journal of Pragmatics*, 37, 685–705.
- Beigman Klebanov, B., & Shamir E. (2005) Guidelines for annotation of concept mention patterns. Technical Report 2005–8, Leibniz Center for Research in Computer Science, The Hebrew University of Jerusalem, Israel.
- Camus, A. (1962). *The stranger*. New York: Vintage Books. Translated from French by Stuart Gilbert. First published in 1942.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249–254.
- Craggs, R., & McGee Wood, M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3), 289–295.
- Di Eugenio, B., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics*, 30(1), 95–101.
- Green, S. (1998) Automated link generation: Can we do better than term repetition? *Computer Networks*, 30(1–7), 75–84.
- Grosz, B., Joshi, A., & Weinstein, S. (1995) Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), 203–225.

²¹ ©Jay Teitel. The extract is reprinted with a kind permission from the author.

- Halliday, M., & Hasan, R. (1976) *Cohesion in English*. London: Longman Group Ltd.
- Hasan R. (1984) Coherence and cohesive harmony. In J. Flood (Eds.), *Understanding reading comprehension* (pp 181–219). International Reading Association, Delaware.
- Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1), 33–64.
- Hirschman, L., Robinson, P. Burger, J. D. & Vilain M. (1998) Automating coreference: The role of annotated training data. CoRR cmp-lg/9803001.
- Hirst, G., & Budanitsky, A. (2005) Correcting real-word spelling errors by restoring lexical cohesion *Natural Language Engineering*, 11(1), 87–111.
- Hoey, M. (1991). *Patterns of Lexis in text*. Oxford University Press: Hong Kong.
- Karamanis, N., Poesio, M. Mellish, C. & Oberlander, J. (2004) Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of ACL-04* (pp. 391–398). Barcelona, Spain
- Krippendorff, K. (1980) *Content analysis*. Beverly Hills, CA: Sage Publications.
- Mann, W., & Thomson S. (1988). Rhetorical structure theory. *Text* 8(3), 243–281.
- Marcu D. (2000) *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge Mass.
- Marcus, M., Santorini, B., Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- Morris, J., & Hirst, G. (1991). Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics*, 17(1), 21–48.
- Morris J., & Hirst, G. (2004) Non-classical lexical semantic relations. In *Proceedings of HLT-NAACL workshop on computational lexical semantics* (pp. 46–51). Boston, MA, USA.
- Morris, J., & Hirst G. (2005) The subjectivity of lexical cohesion in text. In J. C. Chanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text*. Springer, Dodrecht, The Netherlands.
- Passonneau, R. (2004) Computing reliability for coreference annotation. In *Proceedings of LREC* (vol. 4. pp. 1503–1506).
- Poesio, M., & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2), 183–216.
- Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.
- Siddharthan, A., & Copestake A. (2004) Generating referring expressions in open domains. In *Proceedings of ACL-04* pp. 407–414. Barcelona, Spain.
- Siegel S., & Castellan J. N. (1988). *Nonparametric statistics for the behavioral sciences*. Boston, MA: McGraw Hill.
- Silber, G., & McCoy, K. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4), 487–496.
- Stairmand, M. A. (1997) Textual context analysis for information retrieval. In *Proceedings of ACM SIGIR* (pp. 140–147). Philadelphia, PA, USA.
- Stokes N., Carthy J., & Smeaton A. F. (2004). SeLeCT: A lexical cohesion based news story segmentation system. *Journal of AI Communications*, 17(1), 3–12
- Vieira, R., & Poesio, M. (2000) An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4), 539–593.
- Webber, B., & Byron D. K. (Eds.), (2004). *Proceedings of the ACL-2004 workshop on discourse annotation*.