Lexical Cohesion: Some Implications of an Empirical Study

Beata Beigman Klebanov and Eli Shamir

School of Computer Science and Engineering,
The Hebrew University of Jerusalem, 91904, Israel,

Abstract. Lexical cohesion refers to the perceived unity of text achieved by the author's usage of words with related meanings[1]. Data from an experiment with 22 readers aimed at eliciting lexical cohesive patterns they see in 10 texts [2, 3] is used to shed light on a number of theoretical and applied aspects of the phenomenon: which items in the text carry the cohesive load; what are the appropriate data structures to represent cohesive texture; what are the relations employed in cohesive structures.

1 Introduction

The reported study contributes to the research on properties of text, beyond well-formed syntax, that make commonly encountered texts fluent and natural. This endeavor dates back at least to Halliday and Hasan's seminal work on textual cohesion [1]. They identified a number of cohesive constructions: repetition (using the same words, or via repeated reference, substitution and ellipsis), conjunction and lexical cohesion, achieved using word meanings.

Cohesion is a property of a text in the reader's eyes; if a reader is not sensitive to a certain putative structure, it can't make the text cohesive in his eyes. For example, if after presenting people with a text "John went home. He was tired", a researcher asks "Who was tired?" and nobody says "John", this would mean that people are not sensitive to the alleged co-referential cohesive tie between "John" and "he".

As it happens, people do recognize co-reference links. Reader-based checks were performed, often in the framework of creating gold standard for reference resolution software; after investing significant effort in compilation and elaboration of guidelines to annotators, it was possible to achieve relatively good inter-annotator agreements on co-reference links [4, 5].

Co-referential cohesion has been thus shown to have a high degree of reader validity; other types of cohesion are less yielding in this respect. Hasan [6] is especially pessimistic about the possibility of readers' agreement on collocational lexical cohesion, created when words that tend to appear in similar contexts actually appear together, under the condition that there be some recognizable relations between their meanings: "... If someone felt that there is a collocational tie between *dive* and *sea*, on what grounds could such a statement be either rejected or accepted?"[6].

Some pessimism notwithstanding, reader-based exploration of lexical cohesion is an emergent endeavor [7, 2, 3]. In this paper, we briefly describe one such experiment

Beigman Klebanov B., Shamir E. and Pistol I. (2005).
Lexical Cohesion: Some Implications of an Empirical Study.
In *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*, pages 13-21
DOI: 10.5220/0002562900130021

(section 2), and use the experimental data to reflect on theoretical and applied issues in lexical cohesion research (sections 3–5).

2 Annotating Lexical Cohesion Relations: an Experiment

To exemplify the notion of collocational lexical cohesion, Halliday and Hasan [1] mention pairs like *dig/garden*, *ill/doctor*, *laugh/joke*, which strongly remind of the idea of scripts [8]: certain things are expected in certain situations, the paradigm example being menu, tables, waiters and food in a restaurant.

Since some situations may take part in many different scripts – consider a text starting with *Mother died today*[1] – the notion of a script is more helpful in an abductive than in a predictive framework. That is, 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 proceeded with a description of a long illness, one would have known that it can be best explained-by/blamed-upon/abduced-to the previously mentioned lethal outcome. In this case *illness* is said to be **anchored** by *died*, where *illness* is anchored and *died* is its anchor; the anchoring relation is marked *illness→died*.

Beigman Klebanov and Shamir [2, 3] conducted an experiment aimed at eliciting anchoring patterns found by readers in texts. They used 10 texts of different genres, each read by 20 people[2]. Participants first read a 5-page long manual that included an extensive example annotation, as well as short paragraphs highlighting various technical (how to mark multiple and complex anchors) and conceptual issues. People were asked to make an effort to separate personal knowledge from what they think is common knowledge, and general relations from those specifically constructed in the text, using co-reference or predication, i.e. people were discouraged from marking *he→ Jones* just because the text says "Jones was a nice man. He laughed a lot", or *drunkard→Jones* on the basis of "Jones is a drunkard".

For each of the 10 experimental texts, participants were asked to read the text first, and ask themselves, for every item first mentioned in the text[3], 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.

Analyzing the experimental data, Beigman Klebanov [9] identified a subset of data that is considered reliably annotated, using agreement statistics (κ) and a validation experiment, where people were asked to judge anchored-anchor pairs produced in the experiment described above. The reliable subsets contained on average 60% of all items in a text, 36% as anchored and 64% as unanchored. The subsets retain the strongest 25% of all anchors given to the reliably anchored items. Thus, although the subsets thus identified are reliable, in the sense that whatever they contain is robustly found by

[1] This is the first sentence of A. Camus' novel *The Stranger.*

[2] Initially, there were 22 annotators, but 2 were excluded as outliers, following a statistical analysis of the data described in [9].

[3] Verbatim repetition and repetition in an inflected form were considered non-first mentions, and were excluded.

people, they leave out quite a lot of data that possibly reflects inter-personal differences in sensitivity to certain potentially cohesive structures. In the subsequent discussion we will usually use data from the reliable subsets (we call this **core** data), unless stated otherwise.

3 Part of Speech vs. Lexical Cohesion

Describing items that tend to participate in lexical cohesion structures, Halliday and Hasan[1] suggest that "... we can safely ignore repetitive occurrences of fully grammatical (closed system) items like pronouns and prepositions and verbal auxiliaries". Indeed, it turns out that although some annotators did think certain such items participate in anchoring relations, these cases were not systematic across annotators, and the core data contains almost no such items, neither as anchored nor as anchors.

Table 1 shows the distribution of various parts of speech (henceforth, POS) in wordlists for the experimental texts, and the anchoring burden carried by those POS in the whole of the data and in the core subset. We used CLAWS POS tagger for English [10][4] to induce POS tags on the analyzed texts, and corrected manually a small number of mistakes.

Table 1. POS vs. anchoring, averaged across 10 texts. Precision: success rate for marking all members of POS as anchored/anchors; Recall: how much of the anchoring texture is recovered using just members of this POS. *Other* category contains pronouns, conjunctions, verb auxiliaries, prepositions, negation markers, numbers, articles, etc.

POS	Distribution		As Anchored				As Anchor			
	Propor-	Stability	Precision		Recall		Precision		Recall	
	tion	(+/- av.)	All	Core	All	Core	All	Core	All	Core
Noun	32%	16%	95%	48%	39%	64%	88%	39%	51%	64%
Adj+V	30%	50%	86-91%	25-27%	34%	32%	81-91%	23-30%	34%	30%
Adv	10%	50%	67%	3%	8%	2%	65%	9%	5%	2%
ProperN	6%	150%	54%	10%	5%	2%	64%	16%	4%	3%
Other	24%	21%	42%	$< 0.5\%$	13%	$< 0.5\%$	42%	$< 0.5\%$	5%	$< 0.5\%$

Overall, nouns, adjectives and verbs constitute a little less than two thirds of first mentions (62%); they cover 73% of all anchored items and 85% of all anchors; they account for 94-96% of items participating in the core anchoring relations. Thus, a model that concentrates only on nouns, adjectives and verbs has very good recall potential for the core data.

Although above 40% of adverbs, proper nouns and functional categories (grouped under *Other*) are marked as anchored or anchors by some people, only very few adverbs (below 10%) and almost no functional category items make it to the core of consensus. There are somewhat more proper names retained; we note, however, that texts differed greatly in the percentages of proper names (as reflected in stability figures in table 1),

[4] via the free WWW trial service at http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/

from almost none in some literary texts to 16% in one of the news stories. There was much across text variability in anchoring behavior of proper names, too, which could be partially due to the problem of doing statistics on very small numbers.

We note that using just nouns and proper names, as is often done in applied lexical cohesion research [11–13], makes only a moderately successful approximation of anchoring phenomenon: nouns with proper names account for 43-56% of all anchoring relations, and for 66-67% of core ones.

4 Lexical Chains

The anchoring relation described above has a strong affinity to the concept of *lexical chain* - sequence of related words that spans a topical unit of the text [14]. However, we show that the global cohesive structure assigned to a text by patterns of anchoring relations is different from the chain structure.

The most detailed exemplification of lexical chains in text was given by Morris and Hirst [14] (henceforth, M&H). They identified all intuitive lexical chains in a Reader Digest article titled "Outland". The first 12 sentences of this text[5] were used in the experiment by Beigman Klebanov and Shamir [3] to enable a detailed comparison of the two kinds of lexical cohesive structures.

Discussing the choice of candidate words from the text, M&H excluded from consideration closed class words and high frequency words. In the first 12 sentences, 37 different items (we count repetitions including inflections as a single item) were intuitively organized by M&H in 5 groups of sizes {14, 15*, 3, 3*, 3}, with one word appearing in both starred groups; these belong to chains {1, 2, 7, 8, 9}, respectively, in M&H's analysis[6].

Out of the 37 items, 31 (84%) are found in the core of anchoring experiment data. There are, however, 18 items in the core data that were not singled out by M&H, although they are neither closed class items nor very frequent, for example *collective, phone, race, rush, university, windows, school, silence*. If we organize core experimental data in a graph, where an arrow from *b* to *a* means that *a* is an anchor for *b* in the core data, we get 11 disconnected components, of sizes 18, 11, 4, 3, and seven components of size 2. We note that 11 of the 18 extra items are members of 2-item components, for example *school→university*, *sound→silence*, and only 7 are missing from the 36 items in the four largest components of core data, which amounts to 81% coverage. Thus, core experimental data largely accords with M&H's intuitions about which items contribute to the lexical cohesion texture, and thereby gives them stronger experimental backing.

However, the structures assigned to those items differ. Figure 1 shows the two largest connected components from core anchoring data, where a downwards arrow goes from an anchored item to its anchor. Numbers inside the nodes mark the number of chain to which M&H assigned the item; no number means the item was not used by M&H.

[5] The actual text is reproduced in [14], page 36; the article is available online via ACL Anthology: http://acl.ldc.upenn.edu/J/J91/

[6] Chains 3-6 start further down the text.

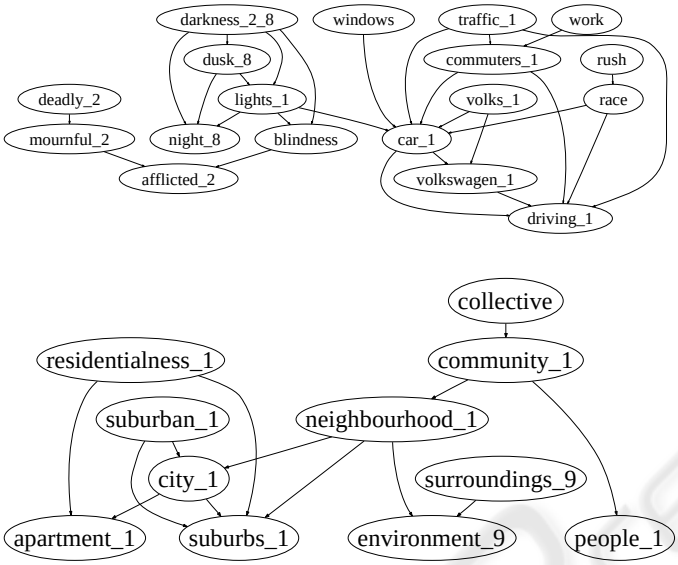

Fig. 1. Anchoring Patterns vs. Lexical Chains

Inspecting the upper component of figure 1, 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 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 ..."[5]. In M&H's analysis, *lights* was assigned to the driving chain (chain 1), and not to any of the other two (chains 2 and 8).

In the second component (bottom half of figure 1) we notice the pivotal position of *neighbourhood*, as a social entity (community, collective, people[7]), as a kind of residential unit (city, suburbs, apartment), and as a physical place (environment, surroundings). The first two aspects were put together by M&H in the same chain as *driving, car*; the third one was identified as a separate chain, but *neighbourhood* was not part of it.

We thus observe that the role assigned to lexical chains - "delineating portions of text that have a strong unity of meaning" (M&H, p.23) - does not use all there is to lexical cohesion. Structures induced from human annotation of elementary relations show a more elaborate picture. Although M&H do not claim that every item should go to just one chain, there is only one case to the contrary in their example (*darkness* is put in chains 2 and 8), and this issue is not explicitly addressed; subsequent research, however, did not allow the same word to participate in multiple chains [11, 12]. Experimental data

[7] *People* is marked with a star in figure 1 because its occurrence in the first 12 sentences was not put in chain 1 by M&H, but a later token of the same item was.

shows that putting every word in at most one chain misses important lexically expressed connections between meaning components that are registered by human readers.

5 Cohesive Semantic Relations

It has been customary in applied lexical cohesion research to use WordNet[15] to determine relatedness ([11, 16, 12, 17, 13]). WordNet is a large lexical database organized by a small number of pre-defined semantic relations, like synonymy, hyponymy, meronymy, etc., termed by Morris and Hirst[18] *classical*, i.e. relations that depend on the sharing of properties, using Lakoff's [19] notion of classical categories.

Since WordNet connects concepts according to only certain kinds of classical relations, it is expected to under-generate when viewed as a lexical cohesion link detector; indeed, Morris and Hirst [18] show that the bulk of cohesive relations are of the non-classical type. This shortcoming could in principle be remedied by including more kinds of relations, or by traversing WordNet not just as a hierarchically structured database, but also as a dictionary, using gloss words, as suggested, for example, in [20].

We would like to raise the question of whether WordNet-style relations over-generate as well. It is possible that some relations thus predicted are not registered by readers (and thus cannot justifiably be said to produce cohesion), because they are overshadowed by other, more salient, relations, not necessarily of an easily classifiable type.

We examine this issue using a text employed both in the experiment by Beigman Klebanov and Shamir [3] and in Barzilay and Elhadad's [21, 11] work on lexical chains for summarization, where WordNet-based chains were identified as a first step, using a typology proposed in [16] - extra strong relations (verbatim repetitions), strong relations (synonym, hypernym, meronym, antonym), and medium strong relations that allowed a path up to the length of four via a common ancestor.

The strongest WordNet-based chain for the 1997 *Economist* article titled "Hello, Dolly"[8] contains various human entities: { adult creator twins parent child sibling son people man dictators master-race slaves tyrant athlete babies progeny victim person }, whereas human readers in Beigman Klebanov and Shamir's anchoring experiment put them in different structures: figures 2 - 5 show the relevant snapshots of the core anchoring relations.

Thus, *creator* belongs to invention, reproduction and divinity; *human, man* are seen from scientific genetic perspective as an organism; *adult, parent, child, sibling, son, babies* form a family related group. We note that these three components are connected - *twins* is the concept mediating between the family branch and the scientific-genetic one; *science, cloned* connect between genetics and divinity. This provides further evidence for the ability of lexical items to participate in more than one meaningful group. The fourth group of humans from Barzilay and Elhadad's list - *tyrant, dictator, slaves* - is in a different component related to the idea of *fear*.

As this example shows, people's appreciation of connections between concepts in the text is finely tuned to the rhetorics of the text, which leave some of the relations

[8] The article and the chains are viewable at:
http://www1.cs.columbia.edu/nlp/summarization-test/summary-test/dolly/textdata1.html

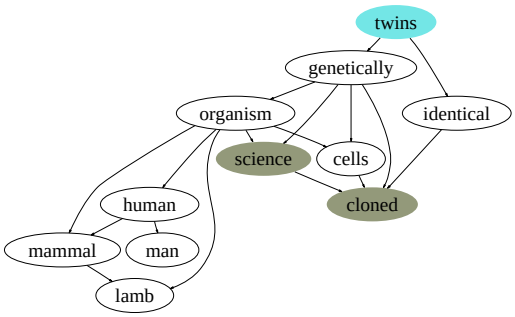

Fig. 2. Organism

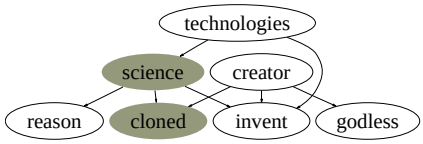

Fig. 3. Creator

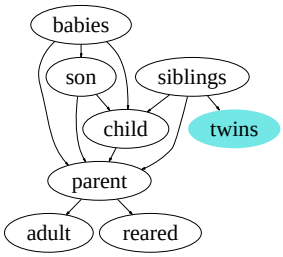

Fig. 4. Family

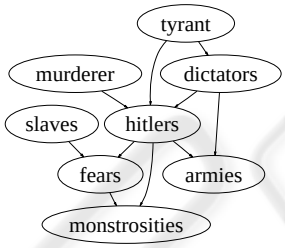

Fig. 5. Tyrant

with clearly identifiable semantics out of the focus. In this example, it seems that discussion of human beings of various kinds is not what makes this text stick together in the readers' eyes.

6 Conclusions

Halliday and Hasan's [1] idea that lexical relations help the text acquire its cohesiveness in the readers' eyes has been subjected to a reader-based test, using the notion of common knowledge based conceptual anchoring [2, 9, 3]. In this paper, the experimental data was used to address a number of theoretical and applied issues in lexical cohesion research.

We showed that nouns, adjectives and verbs carry almost all of the reliably annotated cohesive load; some is left for adverbs and proper names, whereas functional categories are not represented at all. This supports previous suggestions that functional categories are not expected to participate in such relations [1, 14]. However, the assumption usually made in applied models [11–13] that nouns and proper names alone could serve as vehicles of lexical cohesion is not supported, since those cover only about two thirds the data.

We demonstrated that reliance on classical semantic relations for identification of lexical cohesive structure is not entirely justified: although such relations may hold between certain items in the text, people do not necessarily organize the lexical structures in the text according to these relations.

We also argued that representing lexical cohesive patterns by mutually exclusive chains[11–13] undermines rhetorical interconnections between different meaning groups that are sometimes realized lexically, when an item connects back to members of different groups. Thus, a directed graph seems to be a more suitable representation device.

Revealing lexical cohesive structures people see in texts is important from the applied perspective as well. It is expected to improve models of lexical cohesion already employed in applications that analyze human-generated texts: information retrieval [22, 23], text segmentation [13], question answering [24], text summarization [11]. Knowing what humans see there, we are in a better position to guide a machine to look for and make use of the relevant structures.

References

1. Halliday, M., Hasan, R.: Cohesion in English. Longman Group Ltd. (1976)
2. Beigman Klebanov, B., Shamir, E.: Guidelines for annotation of concept mention patterns. Technical Report 2005-8, Leibniz Center for Research in Computer Science, The Hebrew University of Jerusalem, Israel (2005)
3. Beigman Klebanov, B., Shamir, E.: Reader-based exploration of lexical cohesion. Journal paper in preparation (2005)
4. Hirschman, L., Robinson, P., Burger, J.D., Vilain, M.: Automating coreference: The role of annotated training data. CoRR **cmp-lg/9803001** (1998)
5. Poesio, M., Vieira, R.: A corpus-based investigation of definite description use. Computational Linguistics **24** (1998) 183–216
6. Hasan, R.: Coherence and cohesive harmony. In Flood, J., ed.: Understanding Reading Comprehension. Delaware: International Reading Association (1984) 181–219
7. Morris, J., Hirst, G.: The subjectivity of lexical cohesion in text. In Chanahan, J.C., Qu, Y., Wiebe, J., eds.: Computing attitude and affect in text. Springer, Dodrecht, The Netherlands (2005)
8. Schank, R., Abelson, R.: Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum (1977)
9. Beigman Klebanov, B.: Using readers to identify lexical cohesive structures in texts, Workshop submission (2005)
10. Leech, G., Garside, R., Bryant, M.: Claws4: The tagging of the british national corpus. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), Kyoto, Japan (1994) 622–628
11. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: Proceedings of the ACL Intelligent Scalable Text Summarization Workshop. (1997) 86–90
12. Silber, G., McCoy, K.: Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics **28** (2002) 487–496
13. Stokes, N., Carthy, J., Smeaton, A.F.: Select: A lexical cohesion based news story segmentation system. Journal of AI Communications **17** (2004) 3–12
14. Morris, J., Hirst, G.: Lexical cohesion, the thesaurus, and the structure of text. Computational linguistics **17** (1991) 21–48
15. Miller, G.: Wordnet: An on-line lexical database. International Journal of Lexicography **3** (1990) 235–312
16. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., ed.: WordNet: An electronic lexical database. MIT Press, Cambridge, Mass. (1998) 305–332

17. Galley, M., McKeown, K.: Improving word sense disambiguation in lexical chaining. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03). (2003)
18. Morris, J., Hirst, G.: Non-classical lexical semantic relations. In: Proceedings of HLT-NAACL Workshop on Computational Lexical Semantics. (2004)
19. Lakoff, G.: Women, Fire and Dangerous Things. Chicago: University of Chicago Press (1987)
20. Harabagiu, S., Miller, G., Moldovan, D.: Wordnet 2 - a morphologically and semantically enhanced resource. In: Proceedings of SIGLEX-99. (1999) 1–8
21. Barzilay, R.: Lexical chains for summarization. Master's thesis, Ben-Gurion University, Beer-Sheva,Israel (1997)
22. Al-Halimi, R., Kazman, R.: Temporal indexing through lexical chaining. In Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998) 333–351
23. Stairmand, M.A.: Textual context analysis for information retrieval. In: Proceedings of ACM SIGIR. (1997) 140–147
24. Moldovan, D., Novischi, A.: Lexical chains for question answering. In: Proceedings of COLING 2002. (2002)