

Modelling the similarity of discourse connectives

Ben Hutchinson (B.Hutchinson@sms.ed.ac.uk)

School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Abstract

Discourse connectives enable discourse coherence relations to be studied empirically. This paper presents two experiments on the semantic similarity of discourse connectives. Subjects are found to agree significantly on the similarity of pairs of connectives, and their similarity judgements are related to the ability of the connectives to be used to paraphrase each other. Subjects' similarity judgements are also found to correlate positively with the distributional similarity of the connectives.

Introduction

This paper contributes to the empirical study of discourse connectives by considering the problems of measuring and modelling the **similarity** of discourse connectives. The concept of semantic similarity occupies an important role in psychology, artificial intelligence, and computational linguistics. However its applicability to discourse connectives has not been previously studied. If two discourse connectives are found to be similar, this may have implications for theories of the coherence relations that they signal.

Discourse coherence relations contribute to the meaning of texts by specifying relationships between semantic objects such as events and propositions. They also assist in the interpretation of anaphora, verb phrase ellipsis and lexical ambiguities (Hobbs, 1985; Kehler, 2002; Asher & Lascarides, 2003). Some of the many theories of discourse coherence, have been motivated on cognitive grounds. For example, Sanders, Spooren, and Noordman (1992) propose that coherence relations be decomposed into cognitive primitives such as polarity and source of coherence. It is argued that this approach is more likely to be psychologically real than theories which posit relations as indecomposable complex objects.

Knott (1996) argues for the empirical study of the **discourse connectives** (e.g. *because*) that explicitly signal coherence relations, on the basis that relationships between discourse connectives correspond to relationships between discourse coherence relations. Knott argues that if people really do use coherence relations when processing texts, then it is likely that languages will develop ways of signalling these relations explicitly. Discourse connectives thus provide a means of studying coherence relations empirically.

This paper investigates the application of the concept of semantic similarity to discourse connectives. Our mo-

tivation is that knowledge of which connectives are similar can complement theoretical linguistic analysis and so inform theories of discourse coherence. Eliciting similarity ratings for all pairings of discourse connectives is infeasible, however. We therefore explore the hypothesis that the semantic similarity of discourse connectives correlates with their distributional similarity.

Discourse connectives

This section discusses the relationship between discourse connectives and coherence relations. It also introduces Knott's substitution methodology for studying discourse connectives.

Two distinct functions of discourse connectives have been distinguished by Cohen (1984): (1) enabling the faster recognition of coherence relations by the reader, and (2) allowing the recognition of coherence relations which could not be inferred in the absence of a connective. This implies that in some situations the use of a connective by the writer is optional, whereas in others it is required. Moser and Moore (1995) point out that the writer has to decide which connective to use to signal a given coherence relation, as the correspondence between connectives and relations is not one-to-one. For example, both *because* and *seeing as* can be used in (1).

- (1) **Seeing as/because** we've got nothing but circumstantial evidence, it's going to be difficult to get a conviction. (Knott, 1996, p. 177)

This question of whether two discourse connectives can be used to signal the same relation is explored by Knott (1996), who proposes a Test for Substitutability for connectives. The test can be summarised as follows:

1. Take an instance of a discourse connective in a corpus. Imagine you are the writer who produced this text, but that you need to choose an alternative connective.
2. Remove the connective from the text, and insert another connective in its place.
3. If the new connective achieves the same discourse goals as the original one, it is considered **substitutable** in this context.

For example, *because* is substitutable for *seeing as* in (1), but not in (2).

<i>Something happened</i>	despite the fact that	<i>something else happened.</i>
<i>Something happened</i>	even though	<i>something else happened.</i>
<i>(least similar)</i>	○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	<i>(most similar)</i>

Figure 1: Example experimental item

(2) It’s a fairly good piece of work, **seeing as/#because** you have been under a lot of pressure recently. (Knott, p. 177)

By generalising over all the contexts that a connective appears in, Knott defines four possible **substitutability relationships** that can hold between a pair of connectives *X* and *Y*.

- *X* is a **SYNONYM** of *Y* if *X* can always be substituted for *Y*, and vice versa.
- *X* and *Y* are **EXCLUSIVE** if neither can ever be substituted for the other.
- *X* is a **HYPONYM** of *Y* if *Y* can always be substituted for *X*, but not vice versa.
- *X* and *Y* are **CONTINGENTLY SUBSTITUTABLE** if each can sometimes, but not always, be substituted for the other.

Examples of each relationship are given by Knott: *given that* and *seeing as* are **SYNONYMS**, *because* and *seeing as* are **CONTINGENTLY SUBSTITUTABLE**, *on the grounds that* is a **HYPONYM** of *because*, and *because* and *now that* are **EXCLUSIVE**.

Experiment 1:

Eliciting similarity judgements

Semantic similarity is an important concept in cognitive science, but its application to discourse connectives has not previously been studied. Our first hypothesis is that subjects agree on the similarity of connectives.

Hypothesis 1 *Judgements of the similarity of pairs of discourse connectives show significant agreement.*

In semantics, synonymy of nouns or verbs is often defined in terms of the ability to substitute one lexical item for another without affecting the truth of the sentence. Knott’s definition of **SYNONYMY** is closely related to this definition. Based on this relationship between similarity and substitutability, we also make the following two hypotheses.

Hypothesis 2 *Subjects rate pairs of **SYNONYMOUS** connectives as more similar than other pairs of connectives.*

Hypothesis 3 *Subjects rate pairs of **EXCLUSIVE** connectives as less similar than other pairs of connectives.*

Because the substitutability relationships **HYPONYMY** and **CONTINGENTLY EXCLUSIVE** both predict partial inter-substitutability, we do not make predictions regarding the relative similarity of pairs of connectives in these relationships.

Methodology

Materials and design We limit our experiment to discourse connectives which syntactically conjoin clauses or take clausal complements, since adverbial discourse connective have anaphoric properties that complicate interpreting them out of context (Webber, Stone, Joshi, & Knott, 2003). We randomly selected 48 pairs of discourse connectives such that there were 12 pairs standing in each of the four substitutability relationships. To do this, we used substitutability judgements made by Knott (1996), supplemented with judgements of our own. Each experimental item consisted of the two discourse connectives along with the dummy clauses *Something happened* and *something else happened*. An example stimulus item is shown in Figure 1, and the full list of materials is given in the Appendix.

The format of the experimental items was intended to balance two conflicting pressures. Firstly, if discourse connectives are presented on their own, without any sentential context, then it may not always be clear how the item can be used to relate clauses. For example, connectives like *now* and *so* have common uses that are not discourse connectives, and for a connective like *the moment* it may not be obvious to a naive subject that this can connect clauses at all. However, if real example sentences are given to illustrate the connective’s use, then the subject’s judgement may be biased by factors present in those particular example sentences. As a result, the subject may be less likely to consider the full range of situations in which the connective can be used.

We opted for a compromise. We present clausal arguments to each connective, to illustrate how it can be used to relate one clause to another. However the semantic contents of the clauses are left grossly underspecified, so that the subject must imagine for themselves what kind of clauses can be connected in this way. This solution is not perfect, since both clauses are always declarative, and the verb *happen* implies the connective relates events rather than states. Nevertheless, it avoids the problems associated with presenting either a bare lexical item on its own or a completely specified context.

Each subject saw each of the 48 pairs of connectives. The items were presented in a different random order for each subject, and the ordering of the connectives within each item was also randomised.

Procedure Each experiment took approximately 20 minutes. The experiment was conducted remotely over the Internet, with subjects accessing the experiment using their web browser. Data obtained over the web have previously been found to give similar results to data obtained in a laboratory (Keller, 2000).

Relationship	Mean	StdDev	Max	Min
SYNONYMY	3.97	1.33	4.82	3.05
HYPONYMY	3.43	1.51	4.56	1.51
CONT. SUBS.	1.79	1.52	3.10	0.62
EXCLUSIVE	1.08	1.23	2.31	0.55

Table 1: Similarity by substitutability relationship

Instructions Before participating in the experiment, subjects were presented with a set of instructions. The instructions began by explaining that there are words and phrases that can connect sentences, and a number of examples of discourse connectives in context were given. Subjects were then told they would be asked to rate the “similarity in meaning” of pairs of connectives. Three example pairs, illustrating high, medium, and low similarity were given. These were *when-while*, *after-before* and *because-whereas*, respectively. None of these pairs were also used in the experiment. Subjects were explicitly warned that orthographic similarity should not be taken as implying semantic similarity.

After the instructions, subjects completed a short questionnaire. Subjects were asked to provide their name, email, age, sex, handedness and the region where they grew up. Subjects were told if they did not wish to complete the experiment they could submit their partial responses at any time.

Subjects Forty native speakers of English participated in the experiment. Participation was voluntary and unpaid. Of the subjects, 34 were right-handed, 6 left-handed; 15 were female, 25 male. The age of subjects ranged from 21 to 56; the mean was 36 years.

Results and discussion

One subject completed only 16 of the 48 items. Their data are excluded from the calculations of inter-subject agreement, although it is used in the other calculations.

To calculate inter-subject agreement, we used leave-one-out resampling, which is a special case of n -fold cross validation. The average inter-subject correlation was 0.75 (Min = 0.49, Max = 0.86, StdDev = 0.09). A gap could be observed in the mean subjects’ judgements: none of the experimental stimuli had a mean score between 2.37 and 2.84. Instead, the subjects in effect partitioned the pairs of connectives into two bands, representing high and low similarity. These partitions contain 26 and 22 pairs respectively. Average inter-subject correlation within the high similarity partition was 0.42; within the low similarity partition it was 0.45. This shows that the partitioning has a major effect on the overall agreement. Indeed, in only 18% of cases did an individual differ from their peers as to whether a pair of connectives belonged in the high similarity partition or the low similarity partition.

Mean similarity ratings for pairs in each of Knott’s four substitutability relationships are shown in Table 1.

An analysis of variance was conducted, with similarity ratings as the dependent variable. The design had repeated measures of each experimental item, with the human subject (Subj) as a between subject variable, and substitutability relationship (Rel) a within subject variable. Main effects were found for Rel ($F(3, 44) = 40.057, p < 0.001$) and Subj ($F(38, 1672) = 4.767, p < 0.001$). In addition, a crossed effect was found for Subj \times Rel ($F(114, 1672) = 1.963, p < 0.001$), indicating that substitutability affected different subjects’ ratings in different ways. Post-hoc Tukey tests revealed all differences between substitutability relationships to be significant (in each case $p < 0.01$), supporting Hypotheses 2 and 3.

Experiment 2:

Modelling similarity judgements

Given two words, it has been suggested that the more different their contextual distributions are, then also the more semantically different the words will be (Harris, 1970). Conversely, if two words have the same meaning, then they can be expected to have the same contextual distributions. In this experiment we aim to determine whether the distributional similarity of pairs of discourse connectives correlates with the similarity ratings obtained in the previous experiment. To investigate this, we use a lexical co-occurrence model of distributions. In this model, lexical items are treated as co-occurring with a discourse connective if and only if they occur in one of the two clauses related by the connective.

The main verb of a clause introduces the primary predicate, and as such has an important role in determining what coherence relations that clause may be involved in. If a clause contains discourse markers signalling coherence relations (by “discourse markers” we include both structural discourse connectives and discourse adverbials), this can also be expected to contribute to the appropriateness of using a given discourse connective to relate that clause to another. We therefore hypothesise that these features can also be used to predict similarity judgements.

Hypothesis 4 *Discourse connectives with similar verb co-occurrence distributions are rated more similar by subjects than those with dissimilar distributions.*

Hypothesis 5 *Discourse connectives with similar discourse marker co-occurrence distributions are rated more similar by subjects than those with dissimilar distributions.*

Methodology

The subjects’ similarity ratings from the previous experiment were re-used in this experiment. The verb and discourse marker co-occurrences were obtained automatically from a corpus that combined sentences from the British National Corpus with sentences from the internet. One difficulty is that many connectives also have uses where they do not conjoin clauses (e.g. *for* often takes a noun phrase complement). To identify discourse connectives, we first applied an automatic syntac-

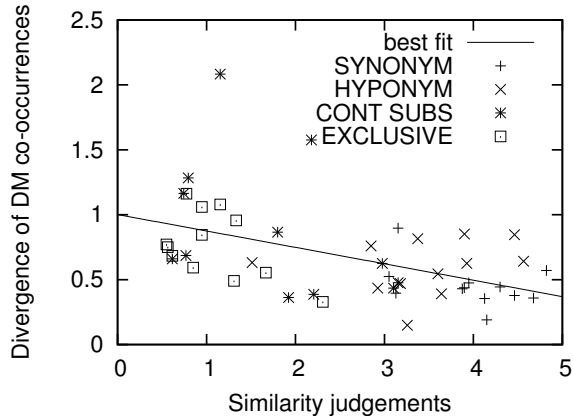


Figure 2: Similarity versus distributional divergence

tic parser (Charniak, 2000) to sentences containing substrings of words matching some connective (and thus potentially contained a connective). The parse trees were then analysed automatically, and if a potential discourse connective occurred in the correct syntactic context then it was positively identified (for details, see (Hutchinson, 2004)). The parse trees enabled us to extract the sets of words occurring in the clauses conjoined by the connective, and provided part of speech information.

Results and discussion

A smoothed variant of the Kullback-Leibler divergence function was used to compare co-occurrence distributions (Lee, 2001). (This function is asymmetric; we applied it with the connectives in alphabetical order.) Spearman’s correlation coefficient for ranked data showed a significant correlation ($r = -0.52$, $p < 0.001$) when context is represented using co-occurrences with discourse markers, but not when context is represented using co-occurrences with verbs. Thus Hypothesis 5 is supported, but Hypothesis 4 is not. This may be because the main predicate of a sentence does not sufficiently constrain what discourse relations a sentence can be an argument of. Conversely, the presence of nearby discourse markers are indicative of a discourse context that the discourse connective must be consistent with.

Figure 2 plots the mean similarity judgements against the distributional divergence obtained using discourse markers, and also shows the substitutability relationship for each item. Two outliers can be observed in the upper left corner; when these are excluded, the magnitude of the correlation drops slightly ($r = -0.51$). The average inter-subject correlation of 0.75 can be considered an upper bound for the task. Recall also that the human subjects effectively partitioned the pairs of connectives into high and low similarity groups. The correlation between distributional divergence (measured using discourse markers) and human judgements within each of these groups is not significant, however.

We also tested whether distributional divergence could

be used to classify pairs of connectives as belonging to the high or low similarity partitions. We divided the pairs of connectives into two groups: those with lower distributional similarity (i.e. high divergence), and those with higher distributional similarity (i.e. low divergence). The boundary between the groups was a KL divergence of 0.6275. This number was chosen so that the group with high distributional similarity had 26 members (the same size as the group which received high mean similarity judgements). Distributional similarity was found to distinguish high vs low similarity judgements with an accuracy of 0.75. Since subjects agreed with their peers 82% of the time as to whether two connectives had a high or low degree of similarity, 0.82 can be considered an upper bound for this task.

Many theories of discourse coherence partition the set of coherence relations through explicit groupings. The fact that distributional similarity correlates with semantic similarity raises the possibility of its informing such groupings. To illustrate how this might be done, Kullback-Leibler divergence scores were calculated between all pairings of the 48 connectives from the previous experiment. These scores were then used to perform agglomerative hierarchical clustering (Manning & Schütze, 1999) of the connectives, and some of the sub-clusters obtained are shown in Figure 3. Numbers indicate the order in which clusters were created, so lower numbers indicate greater similarity. Many of the sub-clusters are linguistically plausible, for example CLUSTER25 and subclusters C8, C12, C16, C11 and C23. The subcluster C7 indicates that *and* and *but* tend to co-occur with the same discourse markers, presumably because each co-occurs with such a wide range of discourse markers.

Related work

Previous studies have shown that subjects show significant agreement when rating the semantic similarity of pairs of nouns or verbs. Rubenstein and Goodenough (1965) presented subjects with 65 pairs of nouns such as *cord-smile* and *gem-jewel* and elicited semantic similarity judgements on a scale of 0–4. The subjects repeated the experiment two weeks later, and the average correlation of each subject’s scores from both sessions was $r = 0.85$. Miller and Charles (1991) elicited similarity judgements for a subset of 30 pairs from Rubenstein and Goodenough’s stimuli. The mean scores they obtained had a correlation of 0.97 with the original mean scores. Resnik (1999) repeated Miller and Charles’ experiment, and calculated an inter-rater agreement of 0.90 by using leave-one-out resampling to compare each subject’s rating with the mean of those of their peers’. Although it has not before been noted, the partitioning effects into high and low similarity groups that we found can also be observed in each of Rubenstein and Goodenough’s, Miller and Charles’ and Resnik’s results. This suggests that this could be a product of the experimental method.

Resnik and Diab (2000) performed a similar experiment with 27 verb pairs (e.g. *bathe-kneel*). In this case, two versions of the stimuli were given: one with the verbs

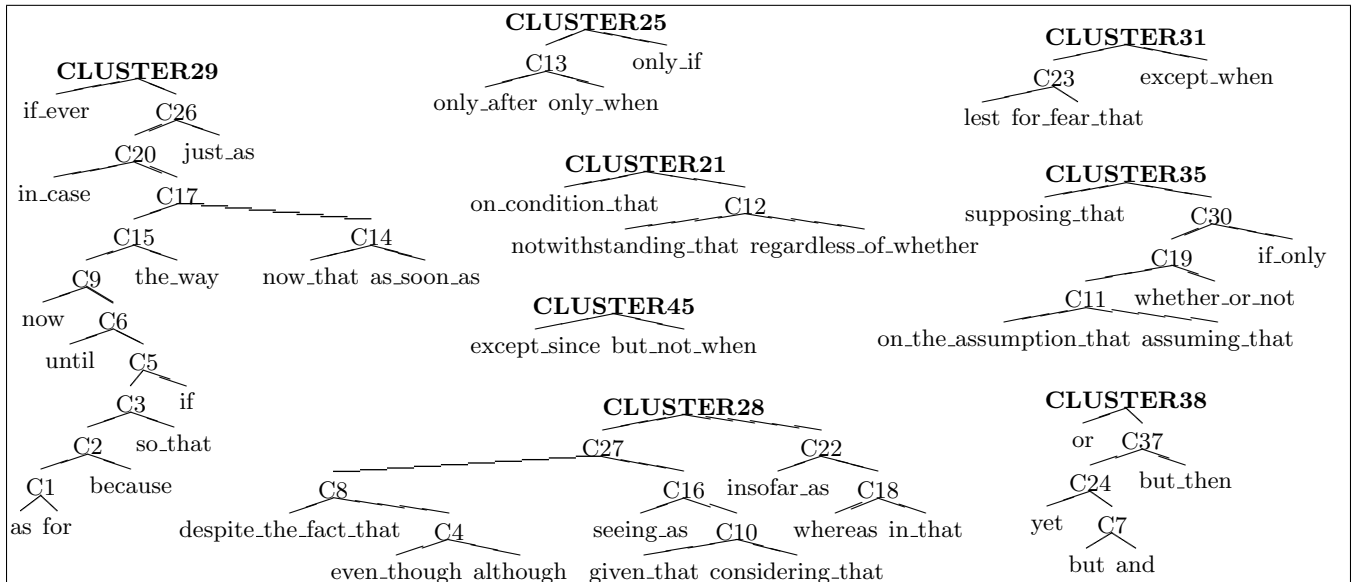


Figure 3: Example clusters of connectives created automatically using distributional similarity

	Inter-subject agreement	Correlation with distributional similarity
Nouns	0.90	0.67
Verbs	0.76	0.43
Discourse Connectives	0.75	0.51

Table 2: Comparison with related work (Resnik, 1999; Resnik & Diab, 2000)

given in a sentential context, the other without context. When context was provided, subjects showed a strong tendency to assign lower similarity ratings in general. In both conditions the level inter-rater agreement was less than that found for nouns: $r = 0.79$ when context was provided; $r = 0.76$ when it wasn't. The difference between conditions may be due to sense disambiguation effects of the contexts.

The studies listed above have also found evidence that similarity ratings correlate positively with the contextual similarity of the lexical items. However the studies differ in how they measure contextual similarity. Miller and Charles use a measure based on sentence completion data, while the other two studies use distributional representations based on lexical co-occurrences. Rubenstein and Goodenough compare the utility of different subsets of co-occurrences for predicting similarity ratings. They find that high frequency content words are much less useful than low frequency content words, while co-occurrences with function words are less useful again. Resnik and Diab use a lexical co-occurrence model that also takes syntactic functions into account.

General Discussion

The experiments extend previous results on the similarity of nouns and verbs. The two main findings are: (1) that humans agree on the similarity of discourse connectives almost as well as they agree on the similarity of nouns, and (2) that human ratings of similarity correlate with the predictions of a distributional model. These findings are remarkable given the complex and abstract nature of the semantics of discourse connectives. Discourse connectives do not have concrete referents, and identifying the relations they signal, let alone defining these relations, can be challenging even for trained analysts. In contrast, for example, almost all the nouns used in previous related studies refer to concrete objects that people are familiar with. People could reasonably be expected to be able to identify the objects that these nouns denote, and even give definitions for them.

Knott (1996) has argued that empirical data on discourse connectives can be used to motivate theories of discourse coherence relations. If we accept this premise, then the similarity of discourse connectives could also be useful in this respect. For example, a theory of coherence which could predict similarity judgements might be considered superior to one that could not. Such an application of gradient linguistic data would have parallels in recent experimental work on grammaticality judgements (Keller, 2000; Sorace & Keller, 2005). Alternatively, distributional similarity might be used to predict substitutability of connectives (Hutchinson, 2005).

Further investigation is required to relate similarity to systems of cognitive primitives proposed to account for coherence relations, such as (Sanders et al., 1992). If two coherence relations have similar decompositions into primitives, then we might expect the discourse connectives that signal those relations (a) to be rated similar by human judges, and (b) to have similar distributions.

References

- Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
- Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the first conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*. Seattle, Washington, USA.
- Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th international conference on computational linguistics* (pp. 251–258).
- Grosz, B. J., & Sidner, C. J. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175–203.
- Harris, Z. S. (1970). *Papers in structural and transformational linguistics*. Dordrecht: Reidel.
- Hobbs, J. A. (1985). *On the coherence and structure of discourse* (Tech. Rep. No. CSLI-85-37). Center for the Study of Language and Information, Stanford University.
- Hutchinson, B. (2004). Mining the web for discourse markers. In *Proceedings of the fourth international conference on language resources and evaluation (lrec 2004)* (pp. 407–410). Lisbon, Portugal.
- Hutchinson, B. (2005). Modelling the substitutability of discourse connectives. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*. Ann Arbor, USA.
- Kehler, A. (2002). *Coherence, reference and the theory of grammar*. Stanford: CSLI publications.
- Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Unpublished doctoral dissertation, University of Edinburgh.
- Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. Unpublished doctoral dissertation, University of Edinburgh.
- Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*, 65–72.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT.
- Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.
- Moser, M., & Moore, J. (1995). Using discourse analysis and automatic text generation to study discourse cue usage. In *Proceedings of the aaai 1995 spring symposium on empirical methods in discourse interpretation and generation* (pp. 92–98).
- Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, 95–130.
- Resnik, P., & Diab, M. (2000, August). Measuring verb similarity. In *Proceedings of the twenty second annual meeting of the cognitive science society (COGSCI2000)*. Philadelphia, US.
- Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Computational Linguistics*, 8, 627–633.
- Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Towards a taxonomy of coherence relations. *Discourse Processes*, 15, 1–35.
- Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, 115(11), 1497–1524.
- Webber, B., Stone, M., Joshi, A., & Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4), 545–588.

Appendix: Materials and mean ratings

SYNONYM pairs	
now–now that (3.13)	although–despite the fact that (4.13)
but–yet (3.90)	considering that–given that (3.88)
or else–or (3.15)	despite the fact that–even though (4.68)
just as–the way (3.05)	considering that–seeing as (4.30)
although–even though (4.15)	regardless of whether–whether or not (4.82)
seeing as–given that (3.95)	on the assumption that–assuming that (4.46)
HYPONYM pairs	
if–if only (2.93)	if–on condition that (4.46)
lest–in case (3.93)	notwithstanding that–even though (3.90)
if–if ever (3.18)	as soon as–the moment (4.56)
and–whereas (1.51)	supposing that–if ever (2.85)
for–because (3.26)	although–notwithstanding that (3.38)
if–assuming that (3.64)	if–on the assumption that (3.60)
CONTINGENTLY SUBSTITUTABLE pairs	
much as [†] –yet (0.79)	but then–much as [†] – (0.74)
given that–in that (3.10)	but–despite the fact that (1.80)
in that–seeing as (2.98)	but not when–by the time (1.15)
if–only if (3.15)	but not when–except since (2.18)
as [‡] –in that (2.21)	for fear that–regardless of whether (0.77)
and–or (0.62)	for–insofar as (1.92)
EXCLUSIVE pairs	
but–only if (0.78)	for fear that–seeing as (0.95)
but–now that (0.95)	just as–supposing that (0.85)
for fear that–until (0.56)	although–except when (1.15)
the way–as [‡] (2.31)	and–assuming that (0.62)
just as–now that (1.67)	only after–whether or not (0.55)
only when–so that (1.31)	considering that–in order that (1.33)

[†] *much as* cannot easily connect events, which may have caused subjects difficulties in rating these items.

[‡] *as* is polysemous (Knott (1996) claims it has three distinct senses). How polysemy affects similarity ratings remains to be explored.