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.

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 motivation 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).
Examples of each relationship are given by Knott: given these relationships.

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 connectives 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).
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 x 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 an noun phrase complement). To identify discourse connectives, we first applied an automatic syntac-

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Mean</th>
<th>StdDev</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYNONYMY</td>
<td>3.97</td>
<td>1.33</td>
<td>4.82</td>
<td>3.05</td>
</tr>
<tr>
<td>HYponymy</td>
<td>3.43</td>
<td>1.51</td>
<td>4.56</td>
<td>1.51</td>
</tr>
<tr>
<td>Cont. Subs.</td>
<td>1.79</td>
<td>1.52</td>
<td>3.10</td>
<td>0.62</td>
</tr>
<tr>
<td>EXCLUSIVE</td>
<td>1.08</td>
<td>1.23</td>
<td>2.31</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 1: Similarity by substitutability relationship
tic parser (Charniak, 2000) to sentences containing sub-
strings of words matching some connective (and thus po-
tentially 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 connect-
ive, and provided part of speech information.

Results and discussion
A smoothed variant of the Kullback-Leibler divergence 
function was used to compare co-occurrence distribu-
tions (Lee, 2001). (This function is asymmetric: we 
applied it with the connectives in alphabetical or-
der.) 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 be-
because the main predicate of a sentence does not suffi-
ciently 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 up-
ner left corner: when these are excluded, the magnitude 
of the correlation drops slightly \( r = -0.51 \). The av-
verage inter-subject correlation of 0.75 can be considered 
an upper bound for the task. Recall also that the hu-
man subjects effectively partitioned the pairs of connect-
ces into high and low similarity groups. The correlation 
between distributional divergence (measured using dis-
course 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 simi-
arity 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 se-
manic similarity raises the possibility of its informing 
such groupings. To illustrate how this might be done, 
Kullback-Leibler divergence scores were calculated be-
tween all pairings of the 48 connectives from the pre-
vious experiment. These scores were then used to per-
form agglomerative hierarchical clustering (Manning & 
Schütze, 1999) of the connectives, and some of the sub-
clusters obtained are shown in Figure 3. Numbers indi-
cate the order in which clusters were created, so lower 
numbers indicate greater similarity. Many of the sub-
clusters are linguistically plausible, for example CLUS-
TER25 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 be-
cause each co-occurs with such a wide range of discourse 
markers.

Related work
Previous studies have shown that subjects show signif-
icant 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 similar-
ity judgements on a scale of 0–4. The subjects repeated 
the experiment two weeks later, and the average corre-
lation 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 rat-
ing 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 experi-
ment with 27 verb pairs (e.g. bathe–kneel). In this case, 
two versions of the stimuli were given: one with the verbs
also takes syntactic functions into account. Resnik and Diab use a lexical co-occurrence model that co-occurrences with function words are less useful again. much less useful than low frequency content words, while ings. They find that high frequency content words are subsets of co-occurrences for predicting similarity rat-

stein and Goodenough compare the utility of different representations based on lexical co-occurrences. Ruben-

fer in how they measure contextual similarity. Miller similarity of the lexical items. However the studies dif-

ferences which could predict similarity judgements might be useful in this respect. For example, a theory of coher-

tance 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 judg-

ments (Keller, 2000; Sorace & Keller, 2005). Alternativa-

ly, 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. Knott (1996) has argued that empirical data on disc-

course 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.

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.

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. Ruben-

stein and Goodenough compare the utility of different subsets of co-occurrences for predicting similarity rat-

ings. 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.

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Inter-subject agreement</th>
<th>Correlation with distributional similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>0.90</td>
<td>0.67</td>
</tr>
<tr>
<td>Verbs</td>
<td>0.76</td>
<td>0.43</td>
</tr>
<tr>
<td>Discourse Connectives</td>
<td>0.75</td>
<td>0.51</td>
</tr>
</tbody>
</table>

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. Ruben-

stein and Goodenough compare the utility of different subsets of co-occurrences for predicting similarity rat-

ings. 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 disc-

course 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 judg-

ments (Keller, 2000; Sorace & Keller, 2005). Alternativa-

ly, 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

Appendix: Materials and mean ratings

<table>
<thead>
<tr>
<th>SYNONYM pairs</th>
<th>HYponYM pairs</th>
<th>CONtINGENTLY SUBSTITUTABLE pairs</th>
<th>EXCLUSIVE pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>now–now that (3.13)</td>
<td>although–despite the fact that (4.13)</td>
<td>much as (0.74)</td>
<td>but–only if (0.78)</td>
</tr>
<tr>
<td>but–yet (3.90)</td>
<td>considering that–given that (3.88)</td>
<td>but then–much as (0.74)</td>
<td>for fear that–seeing as (0.95)</td>
</tr>
<tr>
<td>or else–or (3.15)</td>
<td>despite the fact–even though (4.68)</td>
<td>but–despite the fact (1.80)</td>
<td>just as–the way (0.95)</td>
</tr>
<tr>
<td>just as–the way (3.05)</td>
<td>considering that–seeing as (4.30)</td>
<td>in that–seeing as (2.98)</td>
<td>but not when–by the time (1.15)</td>
</tr>
<tr>
<td>although–even though (4.15)</td>
<td>regardless of whether–or not (4.82)</td>
<td>if–assuming that (3.64)</td>
<td>if–on the assumption that (3.60)</td>
</tr>
<tr>
<td>seeing as–given that (3.95)</td>
<td>on the assumption that–assuming (4.46)</td>
<td>contingent–notwithstanding that (3.38)</td>
<td>for fear that–regardless of whether (0.77)</td>
</tr>
</tbody>
</table>

† 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.