Chapter 5

Empirical evidence for discourse markers at the lexical level

Jelke Bloem
University of Amsterdam

I use a discourse-annotated corpus to demonstrate a new method for identifying potential discourse makers. Discourse markers are often identified manually, but particularly for natural language processing purposes, it is useful to have a more objective, data-driven method of identification. I link this task to the task of identifying co-occurrences of words and constructions, a task where statistical association measures are often used to compute association strengths. I then apply a statistical association measure to the task of discourse marker identification, and present results for several discourse relation types. While the results are noisy due to the limited availability of corpus data, they appear usable after manual correction or as a feature in a classifier. Furthermore, the results highlight a few types of lexical discourse relation cues that are not traditionally considered discourse makers, but still have a clear association with particular discourse relation types.

1 Introduction

The coherence structure of texts is one of many aspects of language that can be studied computationally to gather empirical evidence for previously formulated theories or categorizations. When such theories or categorizations can be automatically mapped to language usage data, this may also provide natural language processing (NLP) systems with more information about a text. Automatically determining the coherence structure of a given text allows for higher-level semantic analysis that is useful in many NLP applications. This structure, consisting of relationships between clauses, is generally described in terms of coherence relations, such as ELABORATION.

There are various frameworks that formalize these relations. One widely used model is Rhetorical Structure Theory (RST) (Mann & Thompson 1988), which provides detailed, tree-shaped structures that cover everything from elementary discourse units to long text spans. At the lowest level of a RST tree structure, discourse relations hold between minimal units of discourse, or spans. These units are generally defined at the level of clauses. Spans can be more central (a nucleus) or secondary to the relation (a satellite). Relations between nuclei and satellites allows us to define the fact that one span fulfills a specific role in the discourse, relative to the other. Multinuclear relations are also possible, such as SEQUENCE. In this case, no single nuclear span can be defined, they are all equally central to the relation. Higher up in a RST tree, these relations also apply to larger spans that cover multiple elementary discourse units. Other examples of formal discourse models are Discourse Representation Theory (DRT) (Kamp 1981; Kamp, van Genabith & Reyle 2011) and the Cognitive Approach to Coherence Relations (CCR) (Sanders, Spooren & Noordman 1992), in which coherence relations consist of features.

Coherence relations in text can be explicitly marked, as in example (1) from the RST Discourse Treebank (Carlson, Okurowski & Marcu 2002), which shows a RESULT relation:

(1) (...) it is better positioned than most companies for the coming overcapacity, because its individual mills can make more than one grade of paper.

Non-lexical markers, such as intonation, are also possible, but not relevant in written text. Coherence relations are often left implicit as well, such as in the following example by Webber (2004):

(2) John is stubborn (C_1). His sister is stubborn (C_2). His parents are stuboorn (C_3). So, they are continually arguing (C_4).

The RESULT relation between C_4 and the other elements is explicitly marked by 'so', but the relations between C_1, C_2 and C_3 are left implicit (in RST, this is a LIST relation).

Explicit lexical markers can be used as cues for systems that attempt to automatically determine discourse relations. There are a limited number of markers known for each kind of discourse relation, so the markers are often identified intuitively and little empirical research has been done to verify their association with the discourse relation(s) they mark. In addition, there might be more markers that are not as obvious, which could be missed in this way.

In this work, I present a method for detecting potential discourse relation markers in a more objective and empirical way. Based on a statistical measure of association, the method used can bring up candidate markers from real-world annotated corpus data purely on the basis of an objective statistical measure. The method will not guarantee a clean list of discourse markers, but it can serve as a first step in detecting them and could have applications as a feature in a larger system for discourse relation detection.

The problem of finding associations between discourse markers and discourse relations can be related to the general issue of discovering whether words co-occur with some other linguistic structure or element. In corpus linguistics, this type of issue is commonly solved using statistical measures of association, such as the χ^2-test. Collocations, words that co-occur more often than would be expected by random chance, are generally analyzed using such methods. Discourse markers and the relations they mark can be viewed from the same perspective — one would expect a discourse marker to be strongly associated with the discourse relation they mark. First I will discuss some related work, and then explain the method and what data is required to apply it to. I will then discuss the results of applying this method to some English-language corpus data, and draw conclusions.

2 Related work

Various studies discuss the identification of discourse markers. Knott (1996) manually examined a set of academic texts to gather a corpus of about 350 discourse markers. To test whether words are discourse markers, he employed a linguistic test. The test involves isolating clauses that contain the markers and checking whether they appear complete that way. If they seem incomplete, then they are considered relational, since they have to be in a relation with another clause to be coherent. This test is argued to be reasonably objective. However, one needs to manually identify candidate discourse markers in advance to be able to perform it, based on intuition. Recent work on discourse connectives in discourse processing also use a pre-defined set of discourse markers, such as Pitler & Nenkova (2009), who use the annotated explicit connectives in the Penn Discourse Treebank and present a method of disambiguating them when they can mark multiple relations.

Timmerman (2007) developed an automatic recognizer for Dutch that relies on discourse markers to generate the RST structure of texts in the medical domain. In his analysis, he produces manually collected lists of Dutch discourse markers, and the relations they signal. He also considers the class of domain-specific markers, in this case medical words, which is interesting since it indicates that (topic) domains are a consideration and that discourse markers are an open class for which new members can be found.

Statistical measure of association have already been applied in discourse structure research, but only for disambiguation purposes. Spenader & Lobanova (2009) have taken two specific relations, CONTRAST and CAUSE-EFFECT, and used the χ^2-test of association strength to determine if intuitively selected markers can reliably distinguish the two. They measure the association between the occurrence of a specific marker in the CONTRAST relation versus the CAUSE-EFFECT relation. However, this is only possible if there is a hypothesis about what relation types a potential marker might belong to, and if only a few relation types are being studied. They surprisingly find that the marker *however*, normally considered a marker of contrast, doesn't help to distinguish CONTRAST and CAUSE-EFFECT. They also find some novel discourse markers. This research shows that statistical methods can result in new findings about discourse markers that intuition does not provide, and do so in a more objective way.

Khazaei, Xiao & Mercer (2015) also present a method of selecting potential lexical coherence markers for RST relations involving the use of n-grams from corpora, to be used in a discourse relation classifier for CIRCUMSTANCE relations. This selection task is similar to the task we describe, but a different method is used. The RST Discourse Treebank (Carlson, Okurowski & Marcu 2002) is used for the extraction of n-grams (up to trigrams). A modified TF-IDF metric is applied to data extracted from the corpus. The relations from the corpus are sorted into documents, one document per relation type, and TF-IDF is then used to select potentially relevant cues that identify a particular document (i.e. relation type). Their study focuses on the CIRCUMSTANCE relation, and results are reported only for this relation. The fact that Khazaei, Xiao & Mercer (2015) use their lexical markers to classify CIRCUMSTANCE relations also shows that automatically detected lexical coherence markers can indeed be used for disource relation classification, although in this work, only one relation type was classified.

3 Method

In this section, I will describe my use of statistical methods of association for detecting potential discourse markers for a given relation type. The basic idea is to find words that occur in a given discourse relation type more often than would be expected by chance, i.e. if the words were randomly distributed over the different relation types. These words can then be considered to be associated with that discourse relation, which indicates that they may be markers for the relation. We also distinguish between nuclei and satellite spans of relations. For example, in a RESULT relation, the thing in the nucleus span is what caused (or might have caused) the thing described in the satellite span. Since different words may be associated with each of these functions, it seems better to distinguish them. However, Khazaei, Xiao & Mercer (2015) do not make this distinction.

This task requires a sufficiently large corpus of discourse-annotated data, from which the relations and the words in their spans should be extracted. One can then calculate the association between these words and the relation nucleus or satellite. Since some markers can consist of multiple words, I included bigrams of the words as well. Trigrams could also be considered, in case there are multi-word potential discourse markers, but this would likely make the results too noisy.

3.1 Use of Context Ratio

The proposed method using measures of statistical association is inspired by the use of association strength in other areas of linguistics. Lichte & Soehn (2007) used the method to discover *negative polarity items* (NPIs). These are lexical constructions that can only occur in the scope of *downward entailing* (DE) operators. By measuring the association strength between words (potential NPIs) and clauses in the scope of DE operators, new NPIs could be found. The task is similar to finding discourse markers, in that there is some class of lexical items that is only partly known, associated with a particular kind of context. The authors of this study used the Context Ratio measure, a very basic measure of association. However, the NPIs that they are looking for are expected to occur in DE contexts most of the time, while discourse markers don't only occur in the discourse relation they mark. For our purposes, it would be better to use a measure that also takes into account how often the marker word appears in other contexts, such as Fisher's Exact Test.

3.2 Use of Fisher's Exact Test

Another example of the use of association strength in linguistics is the method of collostructional analysis, first described by Stefanowitsch & Gries (2003). It is concerned with linguistic constructions, such as [N *waiting to happen*], where N is an open slot that can be filled with a noun. They found that some words are more strongly associated with such a construction than others, providing clues about the construction's meaning. They calculate the association between a word and a construction using Fisher's exact test. This test is particularly well suited to sparse data, a common occurrence in linguistics. Many measures of association make distributional assumptions that are not valid for linguistic data. The occurrence of words in language is not normally distributed, some words occur quite often while many words occur rarely.

Fisher's Exact Test provides a p-value as well as a measure of effect size (Maximum Likelihood Estimate odds ratios), which provides a threshold that makes it possible to state whether an association is statistically significant. It is also an exact computation, while many other association measures are estimations, an important point for this investigation since sparse data is likely. So I have decided to follow Stefanowitsch & Gries (2003) and use Fisher's Exact Test focusing on p-values. However, an empirical validation to determine the optimal association measure for a particular task could show another measure to be better. This was done by Wiechmann (2008) for the task of collostructional analysis. The use of a measure of effect size such as odds ratios is also a viable alternative, when one considers the size of the associations between words and relations to be more important than defining a threshold.

3.3 Use of TF-IDF

The choice for a measure of association, such as Fisher's Exact Test, can be contrasted with Khazaei, Xiao & Mercer's (2015) approach using TF-IDF (Term Frequency - Inverse Document Frequency), which is more commonly used in NLP tasks. This measure is normally used to measure the importance of words in documents, i.e. in the NLP tasks of keyword identification or text summarization. Khazaei, Xiao & Mercer (2015) apply it to the task of discourse marker identification by compiling all spans of a particular relation, such as CIRCUMSTANCE, into a single document, resulting in one document per relation type. Methods that operate on documents, such as keyword identification, can then be employed. The method works by taking the term frequency (TF) (the frequency of a particular word throughout the whole set of documents, i.e. the corpus) and dividing it by the inverse document frequency (IDF). The IDF represents the number of documents that contain the word. In terms of Khazaei, Xiao & Mercer's (2015) study, it represents the number of discourse relation types that contain the word. Therefore, the frequency of a word in the corpus is weighted by the number of different relation types it appears in. Statistically, this is very similar to the way Fisher's Exact Test is computed, however, there is one factor that is involved in the computation of measures of association, but not in computing TF-IDF: the frequency of the word (or term) in a particular relation type. TF-IDF only takes into account whether a term occurs in a relation type, but not how often. It can be argued that important information is lost in this way — a word that occurs only once in a RESULT nucleus is probably less important as a marker of results than a word that occurs in RESULT nuclei many times. Furthermore TF-IDF cannot provide p-values.

4 Data

This section will discuss my source of discourse-annotated data, and show what information should be extracted from it in order to use Fisher's Exact Test to compute association strengths between the discourse relations and the words that are potential markers of these relations. The task requires a sufficiently large corpus of discourse-annotated data, meaning that the text is annotated for discourse units, the relations between them (or at least the low-level relations), and the relation type. Discourse markers don't need to be annotated, since they are what the method is trying to find. I chose to use the RST Discourse Treebank[1] (Carlson, Marcu & Okurowski 2001), a manually annotated treebank with over 178,000 words and over 21,000 discourse units, as my data source. This treebank was also used by Khazaei, Xiao & Mercer (2015) for their discourse marker identification study. While it is smaller than the Penn Discourse Treebank (PDTB), it is based on the widely used Rhetorical Structure Theory (RST) approach, rather than being theory-neutral like the PDTB. The RST annotators claim not to have been influenced by the presence of discourse markers during their analysis (Williams & Reiter 2003). More importantly, the PDTB does not distinguish between nuclei and satellites of relations, an important concept in RST. Instead, relations have a first and a second argument. Discourse markers are more likely to be related to the function of a discourse unit than to its position, so the nucleus-satellite distinction is an important one.

To apply measures of association strength to this data, we gather frequency data from the corpus on discourse types, words that occur in them, and their co-occurrence. p-values are computed from this, where a lower p-value indicates a stronger association. I only consider words in the first, second or last position, or bigrams of the first and second position in case there are two-word markers, because this is where discourse markers tend to appear. For the relations, I only consider the ones between minimal units of discourse. To test all of the discourse relations, I run the test multiple times, once for each relation type.

5 Results

In this section, I will discuss the results of applying this method to the chosen corpus for a few selected types of discourse relation. There are 57 types in the corpus, so I

[1] http://www.isi.edu/~marcu/discourse/Corpora.html.

cannot report on all of them for reasons of space. The following results have been calculated with Fisher's Exact Test, without any frequency cutoffs.

Table 1: Rankings of top discourse marker candidates for nuclei and satellites of Consequence-n relations. The values are the p-values that represent association strength.

because	8.71658e-027
because of	5.62176e-012
of	2.02458e-007
largely because	2.15549e-005
when	2.31501e-005

(a) Consequence-n satellites

and	1.53749e-005
the dollar	1.54064e-005
share	4.21597e-005
loss	9.29785e-005
dollar	9.29785e-005

(b) Consequence-n nuclei

Table 1a shows the five words that are most strongly associated with satellites of consequence-n relations, the same type used in the example contingency table above. *Because* tops the rankings, along with some bigrams involving this word. While I wasn't able to find a list of consequence markers (this relation type is not used in all versions of RST), the definition of *because*, 'for the reason that',[2] seems to imply consequence. The word *of* is not a discourse marker, but is likely listed due to the common collocation 'because of', which also appears in the list and can be considered a multiword discourse marker. The word *when*, appearing 5th in the ranks, tends to occur in this relation when the satellite comes first, in the form 'When x happens, there is a consequence'. Therefore, it can be considered a marker of consequence, even though it doesn't seem to be known as such.

Table 1b shows the top five for nuclei of the same relation type. It contains a stopword and various financial terms, and nothing that could be considered a discourse marker. Yet, the candidate markers that are found still have p-values below common thresholds of statistical significance, such as $\alpha < 0.001$. The financial terms are domain-specific noise, since the RST treebank is based on Wall Street Journal articles. Domain-specific discourse markers also exist (Timmerman 2007), but we do not find them here. It seems that this type of relation is generally marked in the satellite. An example of such a relation helps to illustrate this:

(3) Nuc: *Lockheed reported a $32 million third-quarter net loss,*
Sat: *largely because of cost overruns on fixed-price military contracts.*

The relation here is marked by *largely because* in the satellite, while the nucleus simply reports the consequence (of the cost overruns) without further marking.

The purpose relation is regarded as one of the most commonly marked ones (Taboada 2006). By examining such a relation, there should be less noise among the most strongly associated words. Table 2a shows the result for purpose satellites,

[2] `http://www.merriam-webster.com/dictionary/because`.

Table 2: Rankings of top discourse marker candidates for satellites of Purpose and Concession relations, extracted using the described method.

to	0
to yield	2.31430e-033
yield	3.68036e-032
build	5.59667e-017
to build	7.65295e-014
accommodate	9.69231e-014
to accommodate	9.69231e-014
keep	2.09227e-013

(a) Purpose satellites

even	2.73407e-036
though	4.36469e-030
although	7.37035e-026
despite	1.76170e-021
even if	1.81139e-013
despite the	1.08417e-012
even though	7.25280e-010
if	7.74321-007

(b) Concession satellites

which is where the relation is normally marked. It shows that *to* is very strongly associated with purpose satellites, which is also one of the markers identified by Taboada in spoken dialogue. Most of the other candidate markers in the top part of the list are bigrams of *to* and some verb, and these verbs also occur by themselves in the list (such as *yield*). They are not normally considered discourse markers, but clearly they have a strong association with purpose satellites anyway. Intuitively they do seem to be semantically related, which is actually a claim of the collostructional analysis method that inspired the method used in the present study — one can find out about the meaning of a construction by looking at the words it occurs with. Perhaps the same can be said of discourse relation types.

In the 11th position (not visible in the table) there is also the bigram *in order*, likely used as part of *in order to*, which indicates purpose. Apparently there are some potential discourse markers that will be missed by not including trigrams in the analysis. However, introducing more n-grams also introduces noise. This can be seen in the results of concession satellites in Table 2b, where we find the bigram *despite the* as a potential discourse marker.

In their discourse marker identification study, Khazaei, Xiao & Mercer (2015) focused on the circumstance relation, identifying the following potential lexical cues: *When, after, on, before, with, out, as.* It is difficult to compare these studies directly due to the different choices made in selecting and filtering the data (for example, their study does not distinguish between nucleus and satellite spans), but I will try to provide a comparison here. In my results, the cue *when* can be found as the top ranked cue for circumstance satellites. *as* is ranked second, *after* is ranked fourth. *With* is ranked 15th for satellites, and 40th for nuclei. *On* is ranked 57th for satellites. *Before* is the third ranked cue for circumstance nuclei. *Out* is only ranked as the 264th cue for nuclei, however, *without* is the 6th ranked cue for satellites of circumstance relations. Other cues that are highly ranked in my results but are not mentioned by Khazaei, Xiao & Mercer (2015) include bigrams involving *when*, some high-frequency words such as *the*, and some domain-specific words such as *prices* as a cue for circumstance nuclei. There is clearly some degree of overlap between the results of the two methods, but some of the markers they found were not ranked highly by my method, even though the same corpus was used.

6 Discussion

The results show that statistical association measures can be used to identify potential discourse makers, but that the results are noisy. Taking all of the markers that are statistically significant would result in a far too large list. The significance threshold could be made more strict by applying a correction for performing multiple tests, such as the Bonferroni-correction, but that would only be informative if the data was less noisy. Alternatively, it would be better to use a method of association that is not focused on identifying a particular class of statistically significant discourse markers, but rather on ranking them by their association strength.

The data appears to be noisy due to the use of a large number of domain-specific terms in the corpus, i.e. related to the topics typically discussed in the Wall Street Journal. Khazaei, Xiao & Mercer (2015) deal with domain-specific lexical cues by filtering the lists of candidate discourse markers against another discourse-annotated corpus from a different domain: they remove all cues that fail to succesfully identify the same relations in the secondary corpus. However, this may not always be desireable. It is quite likely that there are domain-specific discourse markers or lexical cues to discourse relations. This didn't come up in the results I presented, but in a previous pilot study with Dutch fundraising letters, a possible example was found: the marker 'fill' for filling in a form occurred in satellites of ENABLEMENT relations. Other domain-specific lexical cues that are strongly associated with the discourse relation, such as the purpose verbs in table 2a, may not necessarily be functional markers of text cohesion, but could still be interesting for natural language processing systems that aim to detect discourse relations in the domain in question. Some of the noise may also come from the small size of the corpus. Perhaps such concerns could be alleviated by combining certain similar discourse relations, or not distinguishing nucleus and satellite spans of relations, but this may make the results less interesting or detailed.

The present results would require further manual processing, for example using Knott's (1996) test, to be useful as lists of actual words with a function of marking coherence. Alternatively, they could be compared to existing, manually compiled lists of discourse markers for that relation, but such lists don't seem to be commonly available, and are made more difficult by the different theories on discourse relations.

Lastly, it would be interesting to investigate why some lexical cues that are not traditionally discourse markers are strongly associated with certain discourse relations. We noted that some of the lexical cues appear to have strong semantic relations to the discourse relation type that they appear in. If speakers use these semantically related elements in interpreting coherence relations, this implies that there might be networks of discourse markers and words with discourse-marking properties. This would provide an interesting parallel to construction grammar theories of language.

Experimental studies could investigate whether speakers of a language can use constructions, rather than just single words, as cues for processing coherence relations between clauses.

Acknowledgements

An older version of this paper was written for John Nerbonne's 2011 Seminar in Methodology and Statistics. I am grateful for his helpful comments on this work.

References

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 1–10.

Carlson, Lynn, Mary Ellen Okurowski & Daniel Marcu. 2002. *RST Discourse Treebank*. Linguistic Data Consortium, University of Pennsylvania.

Kamp, Hans. 1981. A theory of truth and semantic representation. *Formal Methods in the Study of Language*.

Kamp, Hans, Josef van Genabith & Uwe Reyle. 2011. Discourse representation theory. In *Handbook of philosophical logic*, 125–394. Springer.

Khazaei, Taraneh, Lu Xiao & Robert E. Mercer. 2015. Identification and disambiguation of lexical cues of rhetorical relations across different text genres. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, 54.

Knott, A. 1996. *A data-driven methodology for motivating a set of coherence relations*. Department of Artificial Intelligence, University of Edinburgh PhD thesis.

Lichte, T. & J. P. Soehn. 2007. The retrieval and classification of negative polarity items using statistical profiles. *Roots: linguistics in search of its evidential base*. 249–266.

Mann, W. C. & S. A. Thompson. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3). 243–281.

Pitler, Emily & Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 13–16. Association for Computational Linguistics.

Sanders, Ted J. M., Wilbert P. M. Spooren & Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes* 15(1). 1–35.

Spenader, J. & A. Lobanova. 2009. Reliable discourse markers for contrast relations. In *Proceedings of the eighth International Conference on Computational Semantics*, 210–221. Association for Computational Linguistics.

Stefanowitsch, A. & S. T. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Taboada, M. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics* 38(4). 567–592.

Timmerman, S. 2007. *Automatic recognition of structural relations in Dutch text.* MA thesis, University of Twente PhD thesis.

Webber, Bonnie. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28(5). 751–779.

Wiechmann, D. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.

Williams, S. & E. Reiter. 2003. A corpus analysis of discourse relations for natural language generation.