Genre distinctions for Discourse in the Penn TreeBank

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW, UK
bonnie.webber@ed.ac.uk

Abstract

Articles in the Penn TreeBank were identified as being reviews, summaries, letters to the editor, news reportage, corrections, wit and short verse, or quarterly profit reports. All but the latter three were then characterised in terms of features manually annotated in the Penn Discourse TreeBank — discourse connectives and their senses. Summaries turned out to display very different discourse features than the other three genres. Letters also appeared to have some different features. The two main findings involve (1) differences between genres in the senses associated with intra-sentential discourse connectives, inter-sentential discourse connectives and inter-sentential discourse relations that are not lexically marked; and (2) differences within all four genres between the senses of discourse relations not lexically marked and those that are marked. The first finding means that genre should be made a factor in automated sense labelling of non-lexically marked discourse relations. The second means that lexically marked relations provide a poor model for automated sense labelling of relations that are not lexically marked.

1 Introduction

It is well-known that texts differ from each other in a variety of ways, including their *topic*, the *reading level* of their intended audience, and their intended *purpose* (eg, to instruct, to inform, to express an opinion, to summarize, to take issue with or disagree, to correct, to entertain, etc.). This paper considers differences in texts in the well-known Penn TreeBank (hereafter, PTB) and in particular, how these differences show up in the Penn Discourse TreeBank (Prasad et al., 2008).

It first describes ways in which texts can vary (Section 2). It then illustrates the variety of texts to be found in the the PTB and suggests their grouping into four broad *genres* (Section 3). After a brief introduction to the Penn Discourse TreeBank (hereafter, PDTB) in Section 4, Sections 5 and 6 show that these four genres display differences in connective frequency and in terms of the senses associated with intra-sentential connectives (eg, subordinating conjunctions), inter-sentential connectives (eg, inter-sentential coordinating conjunctions) and those inter-sentential relations that are not lexically marked. Section 7 considers recent efforts to induce effective procedures for automated sense labelling of discourse relations that are not lexically marked (Elwell and Baldridge, 2008; Marcu and Echihabi, 2002; Pitler et al., 2009; Wellner and Pustejovsky, 2007; Wellner, 2008). It makes two points. First, because genres differ from each other in the senses associated with such relations, genre should be made a factor in their automated sense labelling. Secondly, because different senses are being conveyed when a relation is lexically marked than when it isn't, lexically marked relations provide a poor model for automated sense labelling of relations that are not lexically marked.

2 Two Perspectives on Genre

The dimension of text variation of interest here is *genre*, which can be viewed *externally*, in terms of the *communicative purpose* of a text (Swales, 1990), or *internally*, in terms of features common to texts sharing a communicative purpose. (Kessler et al., 1997) combine these views by saying that a genre should not be so broad that the texts belonging to it don't share any distinguishing properties —

> ...we would probably not use the term "genre" to describe merely the class of

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 674–682,
Suntec, Singapore, 2-7 August 2009.

> texts that have the objective of persuading someone to do something, since that class – which would include editorials, sermons, prayers, advertisements, and so forth – has no distinguishing formal properties (Kessler et al., 1997, p. 33).

A *balanced* corpus like the Brown Corpus of American English or the British National Corpus, will sample texts from different genres, to give a representative view of how the language is used. For example, the fifteen categories of published material sampled for the Brown Corpus include PRESS REPORTAGE, PRESS EDITORIALS, PRESS REVIEWS and five different types of FICTION.

In contrast, experiments on what genres would be helpful in web search for particular types of information on a topic led (Rosso, 2008), to 18 class labels that his subjects could reliably apply to web pages (here, ones from an .edu domain) with over 50% agreement. These class labels included ARTICLE, COURSE DESCRIPTION, COURSE LIST, DIARY, WEBLOG OR BLOG, FAQ/HELP and FORM. In both Brown's published material and Rosso's web pages, the selected class labels (genres) reflect external purpose rather than distinctive internal features.

Such features are, however, of great interest in both text analysis and text processing. Text analysts have shown that there are indeed interesting features that correlate more strongly with certain genres than with others. For example, (Biber, 1986) considered 41 linguistic features previously mentioned in the literature, including type/token ratio, average word length, and such frequencies as that of particular words (eg, *I/you*, *it*, the proverb *do*), particular word types (eg, place adverbs, hedges), particular parts-of-speech (eg, past tense verbs, adjectives), and particular syntactic constructions (eg, *that*-clauses, *if*-clauses, reduced relative clauses). He found certain clusters of these features (i.e. their presense or absense) correlated well with certain text types. For example, press reportage scored the highest with respect to high frequency of *that*-clauses and contractions, and low type-token ratio (i.e. a varied vocabulary for a given length of text), while general and romantic fiction scored much lower on these features. (Biber, 2003) showed significant differences in the internal structure of noun phrases used in fiction, news, academic writing and face-to-face conversations.

Such features are of similar interest in text processing – in particular, automated genre classification (Dewdney et al., 2001; Finn and Kushmerick, 2006; Kessler et al., 1997; Stamatatos et al., 2000; Wolters and Kirsten, 1999) – which relies on there being reliably detectable features that can be used to distinguish one class from another. This is where the caveat from (Kessler et al., 1997) becomes relevant: A particular genre shouldn't be taken so broadly as to have no distinguishing features, nor so narrowly as to have no general applicability. But this still allows variability in what is taken to be a genre. There is no one "right set".

3 Genre in the Penn TreeBank

Although the files in the Penn TreeBank (PTB) lack any classificatory meta-data, leading the PTB to be treated as a single homogeneous collection of "news articles", researchers who have manually examined it in detail have noted that it includes a variety of "financial reports, general interest stories, business-related news, cultural reviews, editorials and letters to the editor" (Carlson et al., 2002, p. 7).

To date, ignoring this variety hasn't really mattered since the PTB has primarily been used in developing word-level and sentence-level tools for automated language analysis such as wide-coverage part-of-speech taggers, robust parsers and statistical sentence generators. Any genre-related differences in word usage and/or syntax have just meant a wider variety of words and sentences shaping the covereage of these tools. *However, ignoring this variety may actually hinder the development of robust language technology for analysing and/or generating multi-sentence text.* As such, it is worth considering genre in the PTB, since doing so can allow texts from different genres to be weighted differently when tools are being developed.

This is a start on such an undertaking. In lieu of any informative meta-data in the PTB files[1], I looked at line-level patterns in the 2159 files that make up the Penn Discourse TreeBank subset of the PTB, and then manually confirmed the text types I found.[2] The resulting set includes all the

[1] Subsequent to this paper, I discovered that the TIPSTER Collection (LDC Catalog entry LDC93T3B) contains a small amount of meta-data that can be projected onto the PTB files, to refine the semi-automatic, manually-verified analysis done here. This work is now in progress.

[2] Similar patterns can also be found among the 153 files in

genres noted by Carlson et al. (2002) and others as well:

1. Op-Ed pieces and reviews ending with a by-line (73 files): wsj_0071, wsj_0087, wsj_0108, wsj_0186, wsj_0207, wsj_0239, wsj_0257, etc.
2. Sourced articles from another newspaper or magazine (8 files): wsj_1453, wsj_1569, wsj_1623, wsj_1635, wsj_1809, wsj_1970, wsj_2017, wsj_2153
3. Editorials and other reviews, similar to the above, but lacking a by-line or source (11 files): wsj_0039, wsj_0456, wsj_0765, wsj_0794, wsj_0819, wsj_0972, wsj_1259 wsj_1315, etc.
4. Essays on topics commemorating the WSJ's centennial (12 files): wsj_0022, wsj_0339, wsj_0406, wsj_0676, wsj_0933, 2sj_1164, etc.
5. Daily summaries of offerings and pricings in U.S. and non-U.S. capital markets (13 files): wsj_0125, wsj_0271, wsj_0476, wsj_0612, wsj_0704, wsj_1001, wsj_1161, wsj_1312, wsj_1441, etc.
6. Daily summaries of financially significant events, ending with a summary of the day's market figures (14 files): wsj_0178, wsj_0350, wsj_0493, wsj_0675, wsj_1043, wsj_1217, etc.
7. Daily summaries of interest rates (12 files): wsj_0219, wsj_0457, wsj_0602, wsj_0986, etc.
8. Summaries of recent SEC filings (4 files): wsj_0599, wsj_0770, wsj_1156, wsj_1247
9. Weekly market summaries (12 files): wsj_0137, wsj_0231, wsj_0374, wsj_0586, wsj_1015, wsj_1187, wsj_1337, wsj_1505, wsj_1723, etc.
10. Letters to the editor (49 files[3]): wsj_0091, wsj_0094, wsj_0095, wsj_0266, wsj_0268, wsj_0360, wsj_0411, wsj_0433, wsj_0508, wsj_0687, etc.
11. Corrections (24 files): wsj_0104, wsj_0200, wsj_0211, wsj_0410, wsj_0603, wsj_0605, etc.
12. Wit and short verse (14 files): wsj_0139, wsj_0312, wsj_0594, wsj_0403, wsj_0757, etc.
13. Quarterly profit reports – introductory paragraphs alone (11 files): wsj_0190, wsj_0364, wsj_0511, wsj_0696, wsj_1056, wsj_1228, etc.
14. News reports (1902 files)

A complete listing of these classes can be found in an electronic appendix to this article at the PDTB home page (http://www.seas.upenn.edu/~pdtb).

In order to consider discourse-level features distinctive to genres within the PTB, I have ignored, for the time being, both CORRECTIONS and WIT AND SHORT VERSE since they are so obviously different from the other texts, and also QUARTERLY PROFIT REPORTS, since they turn out to be multiple simply copies of the same text because the distinguishing company listings have been omitted.

The remaining eleven classes have been aggregated into four broad genres: ESSAYS (104 files, classes 1-4), SUMMARIES (55 files, classes 5-9), LETTERS (49 files, class 10) and NEWS (1902 files, class 14). The latter corresponds to the Brown Corpus class PRESS REPORTAGE and the class NEWS in the New York Times annotated corpus (Evan Sandhaus, 2008), excluding CORRECTIONS and OBITUARIES. The LETTERS class here corresponds to the NYT class OPINION/LETTERS, while ESSAYS here spans both Brown Corpus classes PRESS REVIEWS and PRESS EDITORIALS, and the NYT corpus classes OPINION/EDITORIALS, OPINION/OPED, FEATURES/XXX/COLUMNS and FEATURES/XXX/REVIEWS, where XXX ranges over Arts, Books, Dining and Wine, Movies, Style, etc. The class called SUMMARIES has no corresponding class in Brown. In the NYT Corpus, it corresponds to those articles whose *taxonomic classifiers* field is NEWS/BUSINESS and whose *types_of_material* field is SCHEDULE.

There are two things to note here. First, no claim is being made that these are the only classes to be found in the PTB. For example, the class labelled NEWS contains a subset of 80 short (1-3 sentence) articles announcing personnel changes – eg, promotions, appointments to supervisory boards, etc. (eg, wsj_0001, wsj_0014, wsj_0066, wsj_0069, wsj_0218, etc.) I have not looked for more specific classes because even classes at this level of specificity show that ignoring genre-specific discourse features can hinder the development of robust language technology for either analysing or generating multi-sentence text. Secondly, no claim is being made that the four selected classes comprise the "right" set of genres for future use of the PTB for discourse-related

the Penn TreeBank that aren't included in the PDTB. However, such files were excluded so that all further analyses could be carried out on the same set of files.

[3]The relation between letters and files is not one-to-one: 13 (26.5%) of these files contain between two and six letters. This is relevant at the end of this section when considering length as a potentially distinguishing feature of a text.

language technology, just that some sensitivity to genre will lead to better performance.

Some simple differences between the four broad genre can be seen in Figure 1, in terms of the average length of a file in words, sentences or paragraphs[4], and the average number of sentences per paragraph. Figure 1 shows that essays are, on average, longer than texts from the other three classes, and have longer paragraphs. The relevance of the latter will become clear in the next section, when I describe PDTB annotation as background for genre differences related to this annotation.

4 The Penn Discourse TreeBank

Genre differences at the level of discourse in the PTB can be seen in the manual annotations of the Penn Discourse TreeBank (Prasad et al., 2008). There are several elements to PDTB annotation. First, the PDTB annotates the arguments of *explicit discourse connectives*:

(1) *Even so, according to Mr. Salmore, the ad was "devastating" because it raised questions about Mr. Courter's credibility*. But **it's building on a long tradition**. (0041)

Here, the explicit connective ("but") is underlined. Its first argument, ARG1, is shown in *italics* and its second, ARG2, in **boldface**. The number 0041 indicates that the example comes from subsection wsj_0041 of the PTB.

Secondly, the PDTB annotates *implicit discourse relations* between adjacent sentences within the same paragraph, where the second does not contain an explicit inter-sentential connective:

(2) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* [Implicit "so"] **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**. (0994)

With implicit discourse relations, annotators were asked to identify one or more explicit connectives that could be inserted to lexicalize the relation between the arguments. Here, they have been identified as the connective "so".

Where annotators could not identify such an *implicit connective*, they were asked if they could identify a non-connective phrase in ARG2 (e.g. "this means") that realised the implicit discourse relation instead (ALTLEX), or a relation holding between the second sentence and an *entity* mentioned in the first (ENTREL), rather than the interpretation of the previous sentence itself:

(3) *Rated triple-A by Moody's and S&P, the issue will be sold through First Boston Corp.* **The issue is backed by a 12% letter of credit from Credit Suisse**.

If the annotators couldn't identify either, they would assert that no discourse relation held between the adjacent sentences (NOREL). Note that because resource limitations meant that *implicit discourse relations* (comprising *implicit connectives*, ALTLEX, ENTREL and NOREL) were only annotated within paragraphs, longer paragraphs (as there were in ESSAYS) could potentially mean more implicit discourse relations were annotated.

The third element of PDTB annotation is that of the senses of connectives, both explicit and implicit. These have been manually annotated using the three-level sense hierarchy described in detail in (Miltsakaki et al., 2008). Briefly, there are four top-level classes:

- TEMPORAL, where the situations described in the arguments are related temporally;
- CONTINGENCY, where the situation described in one argument causally influences that described in the other;
- COMPARISON, used to highlight some prominent difference that holds between the situations described in the two arguments;
- EXPANSION, where one argument expands the situation described in the other and moves the narrative or exposition forward.

TEMPORAL relations can be further specified to ASYNCHRONOUS and SYNCHRONOUS, depending on whether or not the situations described by the arguments are temporally ordered. CONTINGENCY can be further specified to CAUSE and CONDITION, depending on whether or not the existential status of the arguments depends on the connective (i.e. no for CAUSE, and yes for CONDITION).

COMPARISON can be further specified to CONTRAST, where the two arguments share a predicate or property whose difference is being highlighted, and CONCESSION, where "the highlighted differences are related to expectations raised by one

[4] A file usually contains a single article, except (as noted earlier) files in the class LETTERS, which may contain more than one letter.

Genre	Total files	Total paragraphs	Total sentences	Total words	Avg. words per file	Avg. sentences per file	Avg. ¶s per file	Avg. sentences per ¶
ESSAYS	104	1580	4774	98376	945.92	45.9	15.2	3.02
SUMMARIES	55	1047	2118	37604	683.71	38.5	19.1	2.02
LETTERS	49	339	739	15613	318.63	15.1	7.1	2.14
NEWS	1902	18437	40095	837367	440.26	21.1	9.7	2.17

Figure 1: Distribution of Words, Sentences and Paragraphs by Genre (¶ stands for "paragraph".)

argument which are then denied by the other" (Miltsakaki et al., 2008, p.282). Finally, EXPANSION has six subtypese, including CONJUNCTION, where the situation described in ARG2, provides new information related to the situation described in ARG1; RESTATEMENT, where ARG2 restates or redescribes the situation described in ARG1; and ALTERNATIVE, where the two arguments evoke situations taken to be alternatives.

These two levels are sufficient to show significant differences between genres. The only other thing to note is that annotators could be as specific as they chose in annotating the sense of a connective: If they could not decide on the specific type of COMPARISON holding between the two arguments of a connective, or they felt that both subtypes of COMPARISON were being expressed, they could simply sense annotate the connective with the label COMPARISON. I will comment on this in Section 6.

The fourth element of PDTB annotation is attribution (Prasad et al., 2007; Prasad et al., 2008). This was not considered in the current analysis, although here too, genre-related differences are likely.

5 Connective Frequency by Genre

The analysis that follows distinguishes between two kinds of relations associated with explicit connectives in the PDTB: (1) *intra-sentential discourse relations*, which hold between clauses within the same sentence and are associated with subordinating conjunctions, intra-sentential coordinating conjunctions, and discourse adverbials whose arguments occur within the same sentence[5]); and (2) explicit *inter-sentential discourse relations*, which hold across sentences and are associated with explicit *inter-sentential connectives* (inter-sentential coordinating conjunctions and discourse adverbials whose arguments are not in the same sentence).

It is the latter that are effectively in complementary distribution with *implicit discourse relations* in the PDTB[6], and Figures 2 and 3 show their distribution across the four genres.[7] Figure 2 shows that among explicit *inter-sentential connectives*, S-initial coordinating conjunctions ("And", "Or" and "But") are a feature of ESSAYS, SUMMARIES and NEWS but not of LETTERS. LETTERS are written by members of the public, not by the journalists or editors working for the *Wall Street Journal*. This suggests that the use of S-initial coordinating conjunctions is an element of *Wall Street Journal* "house style", as opposed to a common feature of modern writing.

Figure 3 shows several things about the different patterning across genres of *implicit discourse relations* (Columns 4–7 for *implicit connectives*, ALTLEX, ENTREL and NOREL) and *explicit inter-sentential connectives* (Column 3). First, SUMMARIES are distinctive in two ways: While the ratio of implicit connectives to explicit inter-sentential connectives is around 3:1 in the other three genres, for SUMMARIES it is around 4:1 – there are just many fewer explicit inter-sentential connectives. Secondly, while the ratio of ENTREL relations to *implicit connectives* ranges from 0.19 to 0.32 in the other three genres, in SUMMARIES, ENTREL predominates (as in Example 3 from one of the daily summaries of offerings and pricings). In fact, there are nearly as

[5]Limited resources meant that intra-sentential discourse relations associated with subordinators like "in order to" and "so that" or with free adjuncts were not annotated in the PDTB.

[6]This is not quite true for two reasons — first, because the first argument of a discourse adverbial is not restricted to the immediately adjacent sentence and secondly, because a sentence can have both an initial coordinating conjunction and a discourse adverbial, as in "So, for example, he'll eat tofu with fried pork rinds." But it's a reasonable first approximation.

[7]Although annotated in the PDTB, throughout this paper I have ignored the S-medial discourse adverbial *also*, as in "John also eats fish", since such instances are better regarded as presuppositional. That is, as well as a textual antecedent, they can be licensed through inference (e.g. "John claims to be a vegetarian, but he also eats fish.") or accommodated by listeners with respect to the spatio-temporal context (e.g. Watching John dig into a bowl of tofu, one might remark "Don't worry. He also eats fish.") The other discourse adverbials annotated in the PDTB do not have this property.

Genre	Total Sentences	Total Explicit Inter-Sentential Connectives	Density of Explicit Inter-Sentential Connectives/Sentence	S-initial Coordinating Conjunctions	S-initial Discourse Adverbials	S-medial Inter-Sentential Disc Advs
ESSAYS	4774	691	0.145	334 (48.3%)	244 (35.3%)	113 (16.4%)
SUMMARIES	2118	95	0.045	46 (48.4%)	39 (41.1%)	10 (10.5%)
LETTERS	739	85	0.115	26 (30.6%)	37 (43.5%)	18 (21.2%)
NEWS	40095	4709	0.117	2389 (50.7%)	1610 (34.2%)	718 (15.3%)

Figure 2: Distribution of Explicit *Inter-Sentential* Connectives.

Genre	Total Inter-Sentential Discourse Rels	Total Explicit Inter-Sentential Connectives	Implicit Connectives	EntRel	AltLex	NoRel
ESSAYS	3302	691 (20.9%)	2112 (64.0%)	397 (12.0%)	86 (2.6%)	16 (0.5%)
SUMMARIES	916	95 (10.4%)	363 (39.6%)	434 (47.4%)	12 (1.3%)	12 (1.3%)
LETTERS	433	85 (19.6%)	267 (61.7%)	58 (13.4%)	22 (5.1%)	1 (0.2%)
NEWS	23017	4709 (20.5%)	13287 (57.7%)	4293 (18.7%)	504 (2.2%)	224 (1%)

Figure 3: Distribution of *Inter-Sentential Discourse Relations*, including Explicits from Figure 2.

many EntRel relations in *summaries* as the total of explicit and implicit connectives combined.

Finally, it is possible that the higher frequency of alternative lexicalizations of discourse connectives (AltLex) in LETTERS than in the other three genres means that they are not part of *Wall Street Journal* "house style". (Other elements of *WSJ* "house style" – or possibly, news style in general – are observable in the significantly higher frequency of direct and indirect quotations in *news* than in the other three genres. This property is not discussed further here, but is worth investigating in the future.)

With respect to explicit *intra-sentential connectives*, the main point of interest in Figure 4 is that SUMMARIES display a significantly lower density of intra-sentential connectives overall than the other three genres, as well as a significantly lower relative frequency of intra-sentential discourse adverbials. As the next section will show, these intra-sentential connectives, while few, are selected most often to express Contrast and situations changing over time, reflecting the nature of SUMMARIES as regular periodic summaries of a changing world.

6 Connective Sense by Genre

(Pitler et al., 2008) show a difference across Level 1 senses (Comparison, Contingency, Temporal and Expansion) in the PDTB in terms of their tendency to be realised by explicit connectives (a tendency of Comparison and Temporal relations) or by Implicit Connectives (a tendency of Contingency and Expansion). Here I show differences (focussing on Level 2 senses, which are more informative) in their frequency of occurance in the four genres, by type of connective: explicit intra-sentential connectives (Figure 5), explicit inter-sentential connectives (Figure 6), and implicit inter-sentential connectives (Figure 7). SUMMARIES and LETTERS are each distinctly different from ESSAYS and NEWS with respect to each type of connective.

One difference in sense annotation across the four genres harkens back to a comment made in Section 4 – that annotators could be as specific as they chose in annotating the sense of a connective. If they could not decide between specific level n+1 labels for the sense of a connective, they could simply assign it a level n label. It is perhaps suggestive then of the relative complexity of ESSAYS and LETTERS, as compared to NEWS, that the top-level label Comparison was used approximately twice as often in labelling explicit inter-sentential connectives in ESSAYS (7.2%) and LETTERS (9.4%) than in *news* (4.3%). (The top-level labels Expansion, Temporal and Contingency were used far less often, as to be simply noise.) In any case, this aspect of readability may be worth further investigation (Pitler and Nenkova, 2008).

7 Automated Sense Labelling of Discourse Connectives

The focus here is on automated sense labelling of discourse connectives (Elwell and Baldridge, 2008; Marcu and Echihabi, 2002; Pitler et al., 2009; Wellner and Pustejovsky, 2007; Wellner,

Genre	Total Sentences	Total Intra-Sentential Connectives	Density of Intra-Sentential Connectives/Sentence	Subordinating Conjunctions	Intra-Sentential Coordinating Conjunctions	Intra-Sentential Discourse Adverbials
ESSAYS	4774	1397	0.293	808 (57.8%)	438 (31.4%)	151 (10.8%)
SUMMARIES	2118	275	0.130	166 (60.4%)	99 (36.0%)	10 (3.6%)
LETTERS	739	200	0.271	126 (63.0%)	56 (28.0%)	18 (9.0%)
NEWS	40095	9336	0.233	5514 (59.1%)	3015 (32.3%)	807 (8.6%)

Figure 4: Distribution of Explicit *Intra-Sentential* Connectives.

Relation	Essays	Summaries	Letters	News
Expansion.Conjunction	**253 (18.1%)**	50 (18.2%)	31 (15.5%)	**1907 (20.4%)**
Contingency.Cause	208 (14.9%)	37 (13.5%)	**32 (16%)**	1354 (14.5%)
Contingency.Condition	205 (14.7%)	15 (5.5%)	22 (11%)	1082 (11.6%)
Temporal.Asynchronous	187 (13.4%)	54 (19.6%)	19 (9.5%)	1444 (15.5%)
Comparison.Contrast	187 (13.4%)	**56 (20.4%)**	29 (14.5%)	1416 (15.2%)
Temporal.Synchrony	165 (11.8%)	32 (11.6%)	27 (13.5%)	1061 (11.4%)
Total	1397	275	200	9336

Figure 5: Explicit *Intra-Sentential* Connectives: Most common Level 2 Senses

Relation	Essays	Summaries	Letters	News
Comparison.Contrast	**231 (33.4%)**	**47 (49.5%)**	**20 (23.5%)**	**1853 (39.4%)**
Expansion.Conjunction	156 (22.6%)	24 (25.3%)	**20 (23.5%)**	1144 (24.3%)
Comparison.Concession	75 (10.9%)	11 (11.6%)	5 (5.9%)	462 (9.8%)
Comparison	50 (7.2%)	–	8 (9.4%)	204 (4.3%)
Temporal.Asynchronous	40 (5.8%)	1 (1.1%)	5 (5.8%)	265 (5.6%)
Expansion.Instantiation	37 (5.4%)	3 (3.2%)	3 (3.5%)	236 (5.0%)
Contingency.Cause	32 (4.6%)	1 (1.1%)	12 (14.1%)	136 (2.9%)
Expansion.Restatement	27 (3.9%)	–	6 (7.1%)	93 (2.0%)
Total	691	95	85	4709

Figure 6: Explicit *Inter-Sentential* Connectives: Most common Level 2 Senses

Relation	Essays	Summaries	Letters	News
Contingency.Cause	**577 (27.3%)**	70 (19.28%)	**75 (28.1%)**	**3389 (25.5%)**
Expansion.Restatement	395 (18.7%)	62 (17.07%)	55 (20.6%)	2591 (19.5%)
Expansion.Conjunction	362 (17.1%)	**126 (34.7%)**	40 (15.0%)	2908 (21.9%)
Comparison.Contrast	254 (12.0%)	53 (14.60%)	42 (15.7%)	1704 (12.8%)
Expansion.Instantiation	211 (10.0%)	18 (4.96%)	14 (5.2%)	1152 (8.7%)
Temporal.Asynchronous	110 (5.2%)	7 (1.93%)	6 (2.3%)	524 (3.9%)
Total	2112	363	267	13287

Figure 7: *Implicit* Connectives: Most common Level 2 Senses

	Essays			Summaries		
Relation:	Implicit	Inter-Sent	Intra-Sent	Implicit	Inter-Sent	Intra-Sent
Contingency.Cause	**577 (27.3%)**	32 (4.6%)	208 (14.9%)	70 (19.28%)	1 (1.1%)	37 (13.5%)
Expansion.Restatement	395 (18.7%)	27 (3.9%)	4 (0.3%)	62 (17.07%)	–	–
Expansion.Conjunction	362 (17.1%)	156 (22.6%)	**253 (18.1%)**	**126 (34.7%)**	24 (25.3%)	50 (18.2%)
Comparison.Contrast	254 (12.0%)	**231 (33.4%)**	187 (13.4%)	53 (14.60%)	**47 (49.5%)**	**56 (20.4%)**
Expansion.Instantiation	211 (10.0%)	37 (5.4%)	5 (0.3%)	18 (5.0%)	3 (3.2%)	–
Total:	2112	691	1397	363	95	275

Figure 8: Essays and Summaries: Connective sense frequency

	Letters			News		
Relation:	Implicit	Inter-Sent	Intra-Sent	Implicit	Inter-Sent	Intra-Sent
Contingency.Cause	**75 (28.1%)**	12 (14.1%)	**32 (16%)**	**3389 (25.5%)**	136 (2.9%)	1354 (14.5%)
Expansion.Restatement	55 (20.6%)	6 (7.1%)	4 (2%)	2591 (19.5%)	93 (2.0%)	20 (0.2%)
Expansion.Conjunction	40 (15.0%)	**20 (23.5%)**	31 (15.5%)	2908 (21.9%)	1144 (24.3%)	**1907 (20.4%)**
Comparison.Contrast	42 (15.7%)	**20 (23.5%)**	29 (14.5%)	1704 (12.8%)	**1853 (39.4%)**	1416 (15.2%)
Expansion.Instantiation	14 (5.2%)	3 (3.5%)	–	1152 (8.7%)	236 (5.0%)	18 (0.2%)
Total	267	85	200	13287	4709	9336

Figure 9: Letters and News: Connective sense frequency

2008). There are two points to make. First, Figure 7 provides evidence (in terms of differences between genres in the senses associated with inter-sentential discourse relations that are not lexically marked) for taking genre as a factor in automated sense labelling of those relations.

Secondly, Figures 8 and 9 summarize Figures 5, 6 and 7 with respect to the five senses that occur most frequently in the four genre with discourse relations that are not lexically marked, covering between 84% and 91% of those relations. These Figures show that, no matter what genre one considers, different senses tend to be expressed with (explicit) intra-sentential connectives, with explicit inter-sentential connectives and with *implicit connectives*. This means that lexically marked relations provide a poor model for automated sense labelling of relations that are not lexically marked. This is new evidence for the suggestion (Sporleder and Lascarides, 2008) that intrinsic differences between explicit and implicit discourse relations mean that the latter have to be learned independently of the former.

8 Conclusion

This paper has, for the first time, provided genre information about the articles in the Penn TreeBank. It has characterised each genre in terms of features manually annotated in the Penn Discourse TreeBank, and used this to show that genre should be made a factor in automated sense labelling of discourse relations that are not explicitly marked.

There are clearly other potential differences that one might usefully investigate: For example, following (Pitler et al., 2008), one might look at whether connectives with multiple senses occur with only one of those senses (or mainly so) in a particular genre. Or one might investigate how patterns of attribution vary in different genres, since this is relevant to *subjectivity* in text. Other aspects of genre may be even more significant for language technology. For example, whereas the first sentence of a news article might be an effective summary of its contents – e.g.

(4) Singer Bette Midler won a $400,000 federal court jury verdict against Young & Rubicam in a case that threatens a popular advertising industry practice of using "sound-alike" performers to tout products. (wsj_0485)

it might be less so in the case of an essay, even one of about the same length – e.g.

(5) On June 30, a major part of our trade deficit went poof! (wsj_0447)

Of course, to exploit these differences, it is important to be able to automatically identify what genre or genres a text belongs to. Fortunately, there is a growing body of work on genre-based text classification, including (Dewdney et al., 2001; Finn and Kushmerick, 2006; Kessler et al., 1997; Stamatatos et al., 2000; Wolters and Kirsten, 1999). Of particular interest in this regard is whether other news corpora, such as the New York Times Annotated Corpus (Linguistics Data Consortium Catalog Number: LDC2008T19) manifest similar properties to the WSJ in their different genres. If so, then genre-specific extrapolation from the WSJ Corpus may enable better performance on a wider range of corpora.

Acknowledgments

I thank my three anonymous reviewers for their useful comments. Additional thoughtful comments came from Mark Steedman, Alan Lee, Rashmi Prasad and Ani Nenkova.

References

Douglas Biber. 1986. Spoken and written textual dimensions in english. *Language*, 62(2):384–414.

Douglas Biber. 2003. Compressed noun-phrase structures in newspaper discourse. In Jean Aitchison and Diana Lewis, editors, *New Media Language*, pages 169–181. Routledge.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2^{nd} SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, pages 1–8.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identication with connective specic rankers. In *Proceedings of the IEEE Conference on Semantic Computing*.

Evan Sandhaus. 2008. New york times corpus: Corpus overview. Provided with the corpus, LDC catalogue entry LDC2008T19.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35^{th} Annual Meeting of the ACL*, pages 32–38.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the Association for Computational Linguistics*.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the penn discourse treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of COLING*, Manchester.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*, Singapore.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *TAL (Traitement Automatique des Langues)*, 42(2).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Mark Rosso. 2008. User-based identification of web genres. *J American Society for Information Science and Technology*, 59(7):1053–1072.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Text genre detection using common word frequencies. In *Proceedings of the 18^{th} Annual Conference of the ACL*, pages 808–814.

John Swales. 1990. *Genre Analysis*. Cambridge University Press, Cambridge.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments to discourse connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague CZ.

Ben Wellner. 2008. *Sequence Models and Ranking Methods for Discourse Parsing*. Ph.D. thesis, Brandeis University.

Maria Wolters and Mathias Kirsten. 1999. Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the 9^{th} Meeting of the European Chapter of the Assoc. for Computational Linguistics*, pages 142–149, Bergen, Norway.