Genre distinctions and discourse modes: Text types differ in their situation type distributions

Alexis Palmer and Annemarie Friedrich
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
{apalmer,afried}@coli.uni-saarland.de

Abstract

In this paper we explore the relationship between the genre of a text and the types of situations introduced by the clauses of the text, working from the perspective of the theory of discourse modes (Smith, 2003). The typology of situation types distinguishes between, for example, events, states, generic statements, and speech acts. We analyze texts of different genres from two English text corpora, the Penn Discourse TreeBank (PDTB) and the Manually Annotated SubCorpus (MASC) of the Open American National Corpus. Texts of different types – genres in the PDTB and subcorpora in MASC – are segmented into clauses, and each clause is labeled with the type of situation it introduces to the discourse. We then compare the distribution of situation types across different text types, finding systematic differences across genres. Our findings support predictions of the discourse modes theory and offer new insights into the relationship between text types and situation type distributions.

1 Introduction

Language is not a unitary phenomenon, and patterns of language use change according to the type of text under investigation. In natural language processing, furthermore, it has been shown that there are strong effects from both the domain and the genre of texts on the performance of systems performing automatic analysis. These effects are relevant at nearly all levels of analysis, from part-of-speech tagging to discourse parsing, yet they are in some ways poorly understood. For example, there is no single agreed-upon set of text types that suits all levels of analysis, nor are we aware of systematic guidelines for sorting texts into genre categories; this process often relies on human intuition and the claim that "I know [a document of type X] when I see one."

Rather than conceptualizing text type purely as a document-level characteristic, in this study we take inspiration from a theory which targets *text passages* as an intermediate level of representation. The idea is that most texts are in fact a mix of passages of different types. For example, a news story may begin with a short narrative passage which focuses on one individual's reaction to the newsworthy event and then proceed with a more informative discussion of the topic at hand. Smith (2003) identifies five different types of text passages, or **discourse modes**, each of which is associated with certain linguistic characteristics of the text passage. (See Sec. 2 for more on the modes and the linguistic characteristics.) This study investigates how closely the predicted linguistic characteristics of certain text types are reflected in a body of naturally occurring texts.

We focus on genre differences at the level of the clause, considering the types of situations introduced to the discourse by clauses of text. According to Smith, the situation (or **situation entity**) types presented in a text are an important characteristic for distinguishing between the different types of text passages. Using two sets of documents (see Sec. 3) with genre labels, we investigate the distributions of situation types (see Sec. 2.1 for the inventory of situation types) for the different text types. We find systematic differences between news/jokes texts on the one hand and essay/persuasive texts on the other, as the theory predicts. In the final section of the paper, we briefly discuss potential applications of these findings to argumentation mining.

Mode	Distribution of SEs	Progression
NARRATIVE	mostly Event, State	SEs relate to one another; dynamic events advance narrative time
REPORT	mostly Event, State, General Stative	SEs related to Speech Time; time progresses forward & backward from that time
DESCRIPTION	mostly Event, State, ongoing Event	Time is static; text progresses in spatial terms through the scene described
INFORMATION	mostly General Stative	atemporal; progressing on a metaphoric path through the domain of the text
ARGUMENT / COMMENTARY	mostly General Stative, Fact, Proposition	atemporal; progressing on a metaphoric path through the domain of the text

Table 1: Discourse modes and their linguistic correlates according to Smith (2005).

2 Discourse modes: a theory of text passages and their types

Smith (2003) proposes to analyze discourse at the level of the text passage, viewing each individual text as a mixture of text passages. These passages are contiguous regions of text, generally one or more paragraphs, with particular discourse functions. Each passage belongs to one of five discourse modes: NARRATIVE, REPORT, DESCRIPTION, INFORMATION, ARGUMENT/COMMENTARY. Importantly, the modes can be characterized according to two broad classes of linguistic correlates: the mode of progression through the text passage (roughly temporal or atemporal), and the distribution of situation entity types. The modes and their correlates appear in Table 1.

2.1 Situation entities

In this work we are directly concerned with the second type of linguistic correlate: the situation entities. A situation entity (SE) can be thought of as the abstract object introduced to the discourse by a clause of text. The type of the SE introduced by a clause depends on, among other things, the internal temporal properties of the verb and its arguments. The interpretation of the verb constellation may of course by influenced by adverbials and other linguistic factors. We are primarily interested in finite clauses, for the most part assuming that each clause introduces one SE.[1]

The SE types fall into four broad categories. **Eventualities** describe particular situations such as Events (1) or States (2).

(1) The tour guide pointed to the mosaic. (EVENT)

(2) The view from the castle is spectacular. (STATE)

The class of **General Statives** includes Generalizing Sentences (3), which report regularities, and Generic Sentences (4), which make statements about kinds or classes.

(3) Silke often feeds my cats. (GENERALIZING SENTENCE)

(4) The male cardinal has a black beak. (GENERIC SENTENCE)

The third class of SE types are **Abstract Entities**, which differ from the other SE types in how they relate to the world: Eventualities and General Statives are located spatially and temporally in the world, but Abstract Entities are not. Facts (5) are objects of knowledge, and Propositions (6) are objects of belief. In the following examples, the underlined clauses introduce Abstract Entities to the discourse.

(5) I know <u>that his plane arrived at 11:00</u>. (FACT)

(6) I believe <u>that his plane arrived at 11:00</u>. (PROPOSITION)

Finally, we introduce the category **Speech Acts** for clauses whose main function is performative: namely, Questions (7) and Imperatives (8).

(7) Why is it so? (QUESTION)

[1] For a more detailed discussion of situation entities, please see Friedrich and Palmer (2014b). For even more information, see our project page (`http:\\sitent.coli.uni-saarland.de`) and the references cited there, including a detailed annotation manual.

(8) Please sign and return to the sender. (IMPERATIVE)

2.2 Linking situation types and discourse modes: what does the theory predict?

The broad aim of this study is to compare the predictions of the theory to evidence from text corpora, in particular with respect to the distributions of SEs across different text types. We focus on two modes: REPORT and ARGUMENT/COMMENTARY. For the REPORT mode, the expectation is that text passages should be made up primarily of Eventualities (Events and States) with some General Statives. The most frequent SE types in the ARG/COMM mode, on the other hand, should be primarily Abstract Entities (Facts and Propositions) and General Statives.

To date there is no large body of data annotated with discourse modes. Therefore, we instead look directly at the distributions of SEs within text passages for which we have annotated data (Friedrich and Palmer, 2014b), taking the genre category assigned within our text corpora as a proxy for discourse mode. We do this under the assumption that some genres are associated with a certain predominant discourse mode. From that assumption, we consider the average SE distributions per text type to reflect the distributions expected from the predominant mode. Specifically, we map texts from the genres news and jokes to the REPORT mode, and essays and fundraising letters to the ARG/COMM mode.

3 Data for corpus study

We test the predictions of the theory on sets of texts extracted from two different corpora, described below. These corpora were chosen in large part because they both group their texts according to genre. Although the two corpora use a different set of genre labels, both cover the two broad categories we are interested in. Annotation and analysis of the two data sets are described in Sec. 4.

3.1 Penn Discourse TreeBank

The Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) provides annotations of discourse structure over a collection of texts from the Wall Street Journal; these texts are from the Penn TreeBank (Marcus et al., 1993), one of the most widely-used annotated corpora in natural language processing. In addition to discourse structure annotations, PDTB texts are hand-labeled with part-of-speech tags, syntactic structure, and, as of relatively recently, genre designations. Webber (2009) found that the texts in PDTB belong to a number of different categories and, further, that the discourse relations marked in the texts pattern according to the genre of the text. In fact, Webber (2009) inspired the current study, raising the question of whether the SE type distributions found in texts similarly reflect the genre of the text.

PDTB	*news*	790
	essays	1723
MASC	*news*	2563
	jokes	3453
	essays	2404
	letters	1850

Table 2: Number of SE-bearing clauses analyzed per corpus, per genre.

The PDTB texts are predominantly from the *news* genre (roughly 1900 texts), with much smaller numbers of texts from four other genres: *essays* (roughly 170 texts), *letters* (roughly 60 texts), *highlights* (roughly 40 texts), and *errata* (25 texts). From these, we extract 20 news texts and 20 essay texts to be used in our study.

3.2 Manually Annotated Sub-Corpus

The second corpus used in this study is MASC (Ide et al., 2008), the Manually Annotated Sub-Corpus of the Open American National Corpus.[2] Overall, MASC contains roughly 500,000 words of text (both written text and transcribed speech), balanced over 19 text types. In addition to manually-checked annotations of sentence and word boundaries, part-of-speech tags, named entities, and both shallow and deeper syntactic structure, some portions of MASC have been annotated for a number of semantic and pragmatic phenomena. For this study, though, we use only the genre labels and our own SE annotations (see Sec. 4).

For our study, we extract texts from the written part of MASC. We use the texts from four of the genres: *news*, *jokes*, *essays*, and *letters*. The letters fall into two sub-categories (*philanthropic-fundraising* and *solicitation-brochures*), though all of the letters have the same general goal of soliciting donations, whether of money, time, or goods.

[2] http://www.anc.org/data/masc

4 Corpus study

In this section we describe the segmentation and annotation of the data, the situation type inventories reflected in the analysis, and the methodology used for computing results. We then present and discuss our findings.[3]

4.1 Segmentation and annotation

Having selected texts for analysis, we next segmented them into clauses, again following the assumption of one SE per clause (with a few exceptional cases). The PDTB texts were segmented manually by the annotator, and the MASC texts using SPADE (Soricut and Marcu, 2003) with some heuristic post-processing. Each clause was then manually labeled with its SE type.

The PDTB annotations were performed by one paid annotator with extensive background in linguistics, with ample training time but only a minimal annotation manual.

The MASC annotations are part of a large ongoing annotation project with multiple paid annotators, an extensive manual, and a structured training phase. In the latter, we take a feature-driven approach to annotation which improves the quality of the annotations, leading to substantial inter-annotator agreement (see Table 3). In addition to the SE type label, annotators mark each clause with three relevant linguistic features, which are not used in the current study, but which guide the annotators to find the best-fitting SE type label. These are inherent lexical aspect of the verb (Friedrich and Palmer, 2014a), genericity of the main referent, and habituality of the event described. Details regarding the annotation scheme and the benefits of feature-driven annotation appear in Friedrich and Palmer (2014b).

4.2 SE inventories

Each of the two analyses uses a slightly different set of SE types. The main difference between the two is that for the PDTB data annotations were done mostly at a coarse-grained level, and the MASC annotations are more fine-grained.

The PDTB analysis remains close to the inventory of SE types presented in Sec. 2.1, with the modification that three of the four coarse-grained categories (i.e. General Statives, Abstract Entities, and Speech Acts) are treated as SE types. In other words, for each of these categories, we conflate its subtypes into a single higher-level type. States and Events are treated as separate categories. The coarse-grained analysis still captures the relevant distinctions yet allows us to make useful generalizations over the relatively small amount of data.

For MASC, we return to a fine-grained analysis. General Statives and Speech Acts are counted at the fine-grained level, and Abstract Entities do not appear in the analysis at all. We add the REPORT type of situation entity, which is a subtype of EVENTS, designed to capture cases like (9).

(9) ..., said the President of the Squash Association. (REPORT)

genre	clauses	Kappa
news	2563	0.667
jokes	3453	0.756
essays	2404	0.493
letters	1850	0.612

Table 3: Number of clauses, inter-annotator agreement (Cohen's Kappa) for MASC subcorpora.

4.3 Method

For both data sets, we compute the distributions of SE types per genre. For each genre, we collect the counts of situation entity types assigned and then compute the corresponding percentages. For the PDTB data (Figure 2), this is a straightforward analysis, as there was only one annotator.

For MASC (Figure 1), we use the annotations of two annotators to compute the distributions. Annotators are allowed to mark a segment with multiple situation types; we simply use all markings of types to compute the percentages. When annotators disagree, we do not adjudicate but rather count both annotations; when they do agree, we counts two instance of the agreed-upon label. Hence, the statistics presented in Figure 1 present an average over the two annotator's assignments. The distributions shown in Figure 1 all differ significantly ($p < 0.01$) from each other according to a χ^2-test, which means that the SE type distributions of the genres are all significantly different from each other: text types differ in their situation type distributions.

[3] Results from the PDTB portion of the analysis were first presented at the 2009 Texas Linguistics Society conference in Austin, Texas.

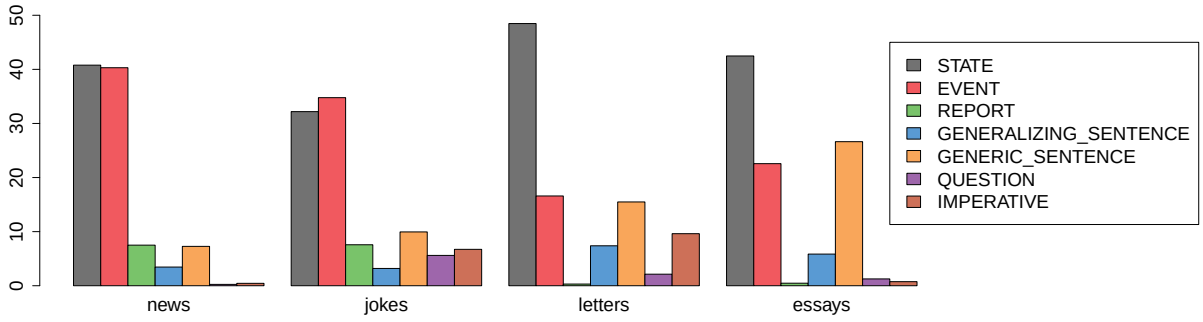

Figure 1: Distributions of situation entity types in four MASC genres.

4.4 Findings

The broad finding is that General Statives play a predominant role for texts associated with the ARGUMENT/COMMENTARY mode, and Events and States for texts associated with the REPORT mode. With these results, we begin to replace the vague distributional statements in Table 1 with more precise characterizations of SE type distributions.

We first compare the two genres shared across both data sets: *news* and *essays*. For both data sets, we see that the proportion of Eventualities is highest for the news genre, and that within Eventualities, Events are more frequent than States.[4] This supports the theoretical claim that passages in REPORT mode predominantly consist of Events and States. Smith (2005) also predicts a significant number of General Statives for REPORT passages; in our study we observe these types in the news texts, but less frequently than Eventualities.[5]

We see more General Statives in essays than in news. The predominance of General Statives is not surprising, given that arguments are frequently built from generalizations and statements about classes or kinds. An interesting result that is not predicted by the theory is that in essays, States are much more frequent than Events. Together with the higher prevelance of General Statives, this suggests that essays rely heavily on describing and discussing states of affairs rather than particular actions or events.

Now we turn to the two additional genres in MASC: *jokes* and *letters*. First it should be noted

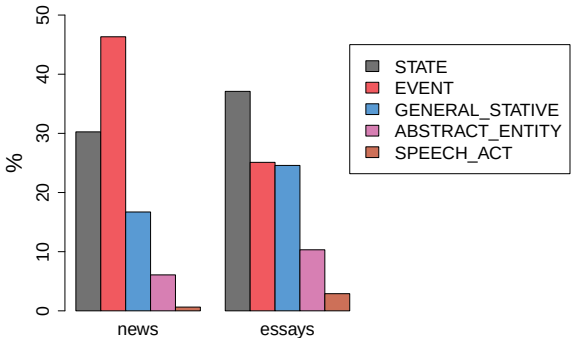

Figure 2: Distributions of situation entity types in two PDTB genres.

that it's not clear whether a distinction should be made between (persuasive) essays and the persuasive letters that appear in MASC. Second, we can see that the predominance of State-type SEs is even stronger for letters than it is for essays. In addition, we see that letters use more generalizing statements and fewer generics, and a rather high proportion of Imperatives. The expected distribution of Imperatives is not explicitly treated by the theory, but one can easily imagine the sorts of Imperative statements that would appear in fundraising and solicitation letters: e.g. "Send a check now! Don't delay! Save the whales!"

Jokes are interesting in that they pattern quite similarly to news texts, but with a higher proportion of Speech Act types. The latter can be attributed to the fact that jokes contain more direct and reported speech than news.

5 Discussion and conclusion

The corpus study described above investigates, across two different datasets of written English text, the relationship between situation entities and text type on the basis of the available data. In

[4]For MASC this second result comes from conflating the categories of Event and Report.

[5]It would be interesting to compare this distribution to texts from another mode (e.g. NARRATIVE) for which Smith (2005) does not predict many General Statives in order to determine the relative importance of General Statives in the REPORT mode.

both cases, and taking genre as a proxy for discourse mode, we find support for Smith's theoretical prediction that different types of text show different characteristic distributions of the types of SEs introduced by the clauses of the text. We find this specifically for two broad text types: news/jokes (mapped to the REPORT mode of discourse) and essays/persuasive texts (mapped to the ARGUMENT/COMMENTARY mode of discourse). The current study analyzes SE distributions over collections of texts; a logical next step is to do this analysis in a more fine-grained fashion, associating SE distributions with text passages labeled with discourse modes. This would remove the need for the genre-as-proxy assumption and move us even further toward a clearer understanding of how discourse modes and situation entity types pattern together.

In future work, we plan to create automatic methods to label clauses with their SE type, which could then be used to automatically identify the types of text passages present in documents.

Relevance for argumentation mining

Some current research in argumentation mining investigates the question of whether performance for automatically extracting argument components from text improves when a system can first narrow down the search space to the argumentative regions of the document. (For example, see Stab and Gurevych (2014) and Levy et al. (2014).) Our finding that essays and persuasive texts show a different distribution of SE types than news texts suggests one way to approach the challenge of finding the argumentative portions of texts.

So far work in argumentation mining has focused predominantly on finding arguments in argumentative texts: opinion pieces, argumentative essays, editorials, and the like. This is to some extent a limiting assumption, as texts from a wide range of genres can in fact contain argumentative passages. A method for finding argumentative passages could extend the range of texts available for argumentation mining.

Acknowledgments

For the PDTB case study, we gratefully acknowledge Caroline Sporleder for much interesting and insightful discussion, as well as Todd Shore both for his annotation work and for discussions arising from that work. This study has also benefitted from discussions with Bonnie Webber and Manfred Pinkal. Finally, huge thanks to the participants of the Bertinoro symposium on the intersection of Argumentation Theory and Natural Language Processing for a highly engaging and intellectually stimulating week. This research was supported in part by the MMCI Cluster of Excellence, and the second author is supported by an IBM PhD Fellowship.

References

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of ACL 2014*.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of The Linguistic Annotation Workshop*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English.

Ran Levy, Yonatan Bilu, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*. Cambridge University Press.

Carlota S Smith. 2005. Aspectual entities and tense in discourse. In *Aspectual Inquiries*, pages 223–237. Springer.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings ACL-HLT 2003*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014*.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of ACL 2009*.