Beyond The Wall Street Journal:
Anchoring and Comparing Discourse Signals across Genres

Yang Liu
Department of Linguistics
Georgetown University
yl879@georgetown.edu

Abstract

Recent research on discourse relations has found that they are cued not only by discourse markers (DMs) but also by other textual signals and that signaling information is indicative of genres. While several corpora exist with discourse relation signaling information such as the Penn Discourse Treebank (PDTB, Prasad et al. 2008) and the Rhetorical Structure Theory Signalling Corpus (RST-SC, Das and Taboada 2018), they both annotate the Wall Street Journal (WSJ) section of the Penn Treebank (PTB, Marcus et al. 1993), which is limited to the news domain. Thus, this paper adapts the signal identification and anchoring scheme (Liu and Zeldes, 2019) to three more genres, examines the distribution of signaling devices across relations and genres, and provides a taxonomy of indicative signals found in this dataset.

1 Introduction

Sentences do not exist in isolation, and the meaning of a text or a conversation is not merely the sum of all the sentences involved: an informative text contains sentences whose meanings are relevant to each other rather than a random sequence of utterances. Moreover, some of the information in texts is not included in any one sentence but in their arrangement. Therefore, a high-level analysis of discourse and document structures is required in order to facilitate effective communication, which could benefit both linguistic research and NLP applications. For instance, an automatic discourse parser that successfully captures how sentences are connected in texts could serve tasks such as information extraction and text summarization.

A discourse is delineated in terms of relevance between textual elements. One of the ways to categorize such relevance is through *coherence*, which

refers to semantic or pragmatic linkages that hold between larger textual units such as CAUSE, CONTRAST, and ELABORATION etc. Moreover, there are certain linguistic devices that systematically signal certain discourse relations: some are generic signals across the board while others are indicative of particular relations in certain contexts. Consider the following example from the Georgetown University Multilayer (GUM) corpus (Zeldes, 2017),[1] in which the two textual units connected by the DM *but* form a CONTRAST relation, meaning that the contents of the two textual units are comparable yet not identical.

(1)   Related cross-cultural studies have resulted in insufficient statistical power, ***but*** interesting trends (e.g., Nedwick, 2014). [academic_implicature]

However, the coordinating conjunction *but* is also a frequent signal of another two relations that can express adversativity: CONCESSION and ANTITHESIS. CONCESSION means that the writer acknowledges the claim presented in one textual unit but still claims the proposition presented in the other discourse unit while ANTITHESIS dismisses the former claim in order to establish or reinforce the latter. In spite of the differences in their pragmatic functions, these three relations can all be frequently signaled by the coordinating conjunction *but*: symmetrical CONTRAST as in (1), CONCESSION as in (2), and ANTITHESIS as in (3). It is clear that *but* is a generic signal here as it does not indicate strong associations with the relations it signals.

(2)   This was a very difficult decision, ***but*** one

---

[1]The square brackets at the end of each example contain the document ID from which this example is extracted. Each ID consists of its genre type and one keyword assigned by the annotator at the beginning of the annotation task.

that was made with the American public in mind. [news_nasa]

(3) NATO had never rescinded it, ***but*** they had and started some remilitarization. [interview_chomsky]

As suggested by Taboada and Lavid (2003), some discourse signals are indicative of certain genres: they presented how to characterize appointment-scheduling dialogues using their rhetorical and thematic patterns as linguistic evidence and suggested that the rhetorical and the thematic analysis of their data can be interpreted functionally as indicative of this type of task-oriented conversation. Furthermore, the study of the classification of discourse signals can serve as valuable evidence to investigate their role in discourse as well as the relations they signal.

One limitation of the RST Signalling Corpus is that no information about the location of signaling devices was provided. As a result, Liu and Zeldes (2019) presented an annotation effort to anchor discourse signals for both elementary and complex units on a small set of documents in RST-SC (see Section 2.2 for details). The present study addresses methodological limitations in the annotation process as well as annotating more data in more genres in order to investigate the distribution of signals across relations and genres and to provide both quantitative and qualitative analyses on signal tokens.

2 Background

Rhetorical Structure Theory (RST, Mann and Thompson 1988) is a well-known theoretical framework that extensively investigates discourse relations and is adopted by Das and Taboada (2017) and the present study. RST is a functional theory of text organization that identifies hierarchical structure in text. The original goals of RST were discourse analysis and proposing a model for text generation; however, due to its popularity, it has been applied to several other areas such as theoretical linguistics, psycholinguistics, and computational linguistics (Taboada and Mann, 2006).

RST identifies hierarchical structure and nuclearity in text, which categorizes relations into two structural types: NUCLEUS-SATELLITE and MULTINUCLEAR. The NUCLEUS-SATELLITE structure reflects a hypotactic relation whereas the MULTINUCLEAR structure is a paratactic relation

(Taboada and Das, 2013). The inventory of relations used in the RST framework varies widely, and therefore the number of relations in an RST taxonomy is not fixed. The original set of relations defined by Mann and Thompson (1988) included 23 relations. Moreover, RST identifies textual units as Elementary Discourse Units (EDUs), which are non-overlapping, contiguous spans of text that relate to other EDUs (Zeldes, 2017). EDUs can also form hierarchical groups known as complex discourse units.

2.1 Relation Signaling

When it comes to relation signaling, the first question to ask is what a signal is. In general, signals are the means by which humans identify the realization of discourse relations. The most typical signal type is DMs (e.g. 'although') as they provide explicit and direct linking information between clauses and sentences. As mentioned in Section 1, the lexicalized discourse relation annotations in PDTB have led to the discovery of a wide range of expressions called ALTERNATIVE LEXICALIZATIONS (*AltLex*) (Prasad et al., 2010). RST-SC provides a hierarchical taxonomy of discourse signals beyond DMs (see Figure 1 for an illustration, reproduced from Das and Taboada (2017, p.752).

Intuitively, DMs are the most obvious linguistic means of signaling discourse relations, and therefore extensive research has been done on DMs. Nevertheless, focusing merely on DMs is inadequate as they can only account for a small number of relations in discourse. To be specific, Das and Taboada (2017) reported that among all the 19,847 signaled relations (92.74%) in RST-SC (i.e. 385 documents and all 21,400 annotated relations), relations exclusively signaled by DMs only account for 10.65% whereas 74.54% of the relations are exclusively signaled by other signals, corresponding to the types they proposed.

2.2 The Signal Anchoring Mechanism

As mentioned in Section 1, RST-SC does not provide information about the location of discourse signals. Thus, Liu and Zeldes (2019) presented an annotation effort to anchor signal tokens in the text, with six categories being annotated. Their results showed that with 11 documents and 4,732 tokens, 923 instances of signal types/subtypes were anchored in the text, which accounted for over 92% of discourse signals, with the signal type *se-*
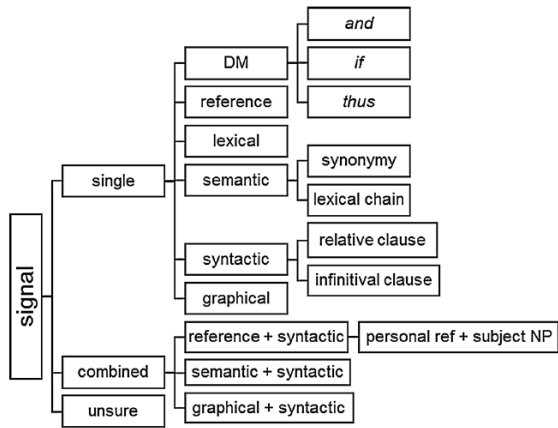
Figure 1: Taxonomy of Signals in RST-SC (Fragment).

Figure 2: A Visualization of How Strongly Each Genre Signals in the GUM Corpus.

mantic representing the most cases (41.7% of signaling anchors) whereas discourse relations anchored by DMs were only about 8.5% of anchor tokens in this study, unveiling the value of signal identification and anchoring.

2.3 Neural Modeling for Signal Detection

Zeldes (2018a) trained a Recurrent Neural Network (RNN) model for the task of relation classification, and then latent associations in the network were inspected to detect signals. It is relatively easy to capture DMs such as 'then' or a relative pronoun 'which' signaling an ELABORATION. The challenge is to figure out what features the network needs to know about beyond just word forms such as meaningful repetitions and variable syntactic constructions. With the human annotated data from the current project, it is hoped that more insights into these aspects can help us engineer meaningful features in order to build a more informative computational model.

3 Methodology

Corpus. The main goal of this project is to anchor and compare discourse signals across genres, which makes the Georgetown University Multilayer (GUM) corpus the optimal candidate, in that it consists of eight genres including interviews, news stories, travel guides, how-to guides, academic papers, biographies, fiction, and forum discussions. Each document is annotated with different annotation layers including but not limited to dependency (`dep`), coreference (`ref`), and rhetorical structures (`rst`). For the purpose of this study, the `rst` layer is used as it includes annotation on discourse relations, and signaling infor-
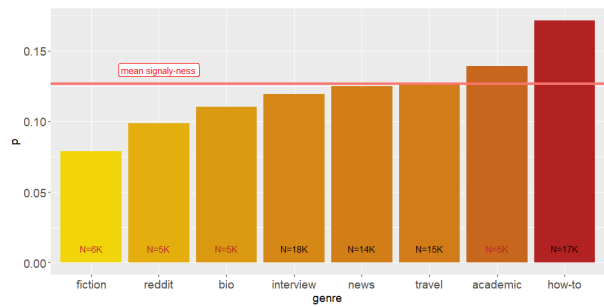
mation will be anchored to it in order to produce a new layer of annotation. However, it is worth noting that other annotation layers are great resources to delve into discourse signals on other levels.

Moreover, due to time limitations and the fact that this is the first attempt to apply the taxonomy of signals and the annotation scheme to other genres outside RST-DT's newswire texts, four out of eight genres in the GUM corpus were selected: *academic*, *how-to guides*, *interviews*, and *news*, which include a collection of 12 documents annotated for discourse relations. The rationale for choosing these genres is that according to Zeldes (2018a)'s neural approach to discourse signal prediction on the GUM corpus, how-to guides and academic articles in the GUM corpus signal most strongly, with interviews and news articles slightly below the average and fiction and reddit texts the least signaled, as shown in Figure 2 (reproduced from Zeldes (2018b, p.19)). It is believed that the selection of these four genres is a good starting point of the topic under discussion.

Annotation Tool. One of the reasons that caused low inter-annotator agreement (IAA) in Liu and Zeldes (2019) is the inefficient and error prone annotation tools they used: no designated tools were available for the signal anchoring task at the time. We therefore developed a better tool tailored to the purpose of the annotation task. It is built over an interface offering full RST editing capabilities called rstWeb (Zeldes, 2016) and provides mechanisms for viewing and editing signals (Gessler et al., 2019).

Annotation Reliability. In order to evaluate the reliability of the scheme, a revised inter-annotator agreement study was conducted using the same metric and with the new interface on three documents from RST-SC, containing 506 tokens with just over 90 signals. Specifically, agreement is

measured based on token spans. That is, for each token, whether the two annotators agree it is signaled or not. The results demonstrate an improvement in Kappa, 0.77 as opposed to the previous Kappa 0.52 in Liu and Zeldes (2019).

Taxonomy of Discourse Signals. The most crucial task in signaling annotation is the selection of signal types. The taxonomy of discourse signals used in this project is adapted from that of Das and Taboada (2017), with additional types and subtypes to better suit other genres. Two new types and four new subtypes of the existing types are proposed: the two new types are *Visual* and *Textual* in which the subtype of the former is *Image* and the subtypes of the latter are *Title, Date,* and *Attribution*. The three new subtypes are *Modality* under the type *Morphological* and *Academic article layout, Interview layout* and *Instructional text layout* under the type *Genre*.

Signal Anchoring Example. Semantic features have several subtypes, with *lexical chain* being the most common one. Lexical chains are annotated for words with the same lemma or words or phrases that are semantically related. Another characteristic of lexical chains is that words or phrases annotated as lexical chains are open to different syntactic categories. For instance, the following example shows that the relation RESTATEMENT is signaled by a *lexical chain* item corresponding to the phrase *a lot of* in the nucleus span and *quantity* in the satellite span respectively.

(4)   [They compensate for this by creating the impression that they have ***a lot of*** friends –]$_N$ [they have a '***quantity***', not quality' mentality.]$_S$ [whow_arrogant]

4 Results & Analysis

This pilot study annotated 12 documents with 11,145 tokens across four different genres selected from the GUM corpus. *Academic articles, how-to guides*, and *news* are written texts while *interview* is spoken language. Generally speaking, all 20 relations used in the GUM corpus are signaled and anchored. However, this does not mean that all occurrences of these relations are signaled and anchored. There are several signaled but unanchored relations, as shown in Table 1. In particular, the 5 unsignaled instances of the relation JOINT result from the design of the annotation scheme (see Section 5.1 for details). Additionally, the unanchored signal types and subtypes are usually associated with high-level discourse relations and usually correspond to genre features such as *interview layout* in interviews where the conversation is constructed as a question-answer scheme and thus rarely anchored to tokens.

With regard to the distribution of the signal types found in these 12 documents, the 16 distinct signal types amounted to 1263 signal instances, as shown in Table 2. There are only 204 instances of DMs out of all 1263 annotated signal instances (16.15%) as opposed to 1059 instances (83.85%) of other signal types. In RST-SC, DM accounts for 13.34% of the annotated signal instances as opposed to 81.36%[2] of other signal types (Das and Taboada, 2017). The last column in Table 2 shows how the distribution of each signal type found in this dataset compares to RST-SC. The reason why the last column does not sum to 100% is that not all the signal types found in RST-SC are present in this study such as the combined signal type *Graphical + syntactic*. And since *Textual* and *Visual* are first proposed in this study, no results can be found in RST-SC, and the category *Unsure* used in RST-SC is excluded from this project.

4.1 Distribution of Signals across Relations

Table 3 provides the distribution of discourse signals regarding the relations they signal. The first column lists all the relations used in the GUM corpus. The second column shows the number of signal instances associated with each relation. The third and fourth columns list the most signaled and anchored type and subtype respectively.

The results show a very strong dichotomy of relations signaled by DMs and semantic-related signals: while DMs are the most frequent signals for five of the relations – CONDITION, CONCESSION, ANTITHESIS, CAUSE, and CIRCUMSTANCE, the rest of the relations are all most frequently signaled by the type *Semantic* or *Lexical*, which, broadly speaking, are all associated with open-class words as opposed to functional words or phrases. Furthermore, the type *Lexical* and its subtype *indicative word* seem to be indicative of JUSTIFY and EVALUATION. This makes sense due to the nature of the relations, which requires writers' or speakers' opinions or inclinations for the subject under discussion, which are usually expressed through positive or negative adjectives (e.g. *serious, outstanding, disappointed*) and other

---

[2]This result excludes the class *Unsure* used in RST-SC.

unanchored relations	frequency	percentage (%)
PREPARATION	28	22.2
SOLUTIONHOOD	11	32.35
JOINT	5	1.92
BACKGROUND	3	2.68
CAUSE	1	4
EVIDENCE	1	4.2
MOTIVATION	1	4.76

Table 1: Distribution of Unanchored Relations.

signal_type	frequency	percentage (%)	RST-SC (%)
Semantic	563	44.58	24.8
DM	204	16.15	13.34
Lexical	156	12.35	3.89
Reference	71	5.62	2.00
Semantic + syntactic	51	4.04	7.36
Graphical	46	3.64	3.46
Syntactic	44	3.48	29.77
Genre	30	2.38	3.22
Morphological	26	2.06	1.07
Syntactic + semantic	25	1.98	1.40
Textual	24	1.90	N/A
Numerical	8	0.63	0.09
Visual	7	0.55	N/A
Reference + syntactic	3	0.24	1.86
Lexical + syntactic	3	0.24	0.41
Syntactic + positional	2	0.16	0.23
Total	1263	100.00	92.9

Table 2: Distribution of Signal Types and its Comparison to RST-SC.

syntactic categories such as nouns/noun phrases (e.g. *legacy, excitement, an unending war*) and verb phrases (e.g. *make sure, stand for*). Likewise, words like *Tips, Steps*, and *Warnings* are indicative items to address communicative needs, which is specific to a genre, in this case, the how-to guides. It is also worth pointing out that EVALUATION is the only discourse relation that is not signaled by any DMs in this dataset.

Even though some relations are frequently signaled by DMs such as CONDITION and ANTITHESIS, most of the signals are highly lexicalized and indicative of the relations they indicate. For instance, signal tokens associated with the relation RESTATEMENT tend to be the repetition or paraphrase of the token(s). Likewise, most of the tokens associated with EVALUATION are strong positive or negative expressions. As for SEQUENCE, in addition to the indicative tokens such as *First & Second* and temporal expressions such as *later*, an indicative word pair such as *stop & update* can also suggest sequential relationship. More inter-

signaled relations	signal instances	signal type	signal subtype
JOINT	260	Semantic (147)	lexical chain (96)
ELABORATION	243	Semantic (140)	lexical chain (96)
PREPARATION	129	Semantic (54)	lexical chain (30)
BACKGROUND	112	Semantic (62)	lexical chain (42)
CONTRAST	68	Semantic (39)	lexical chain (31)
RESTATEMENT	60	Semantic (34)	lexical chain (28)
CONCESSION	49	DM (23)	DM (23)
JUSTIFY	49	Lexical (25)	indicative word (23)
EVALUATION	42	Lexical (31)	indicative word (31)
SOLUTIONHOOD	34	Semantic (12)	lexical chain (5)
CONDITION	31	DM (25)	DM (25)
ANTITHESIS	31	DM (12)	DM (12)
SEQUENCE	26	Semantic (7)	lexical chain (6)
CAUSE	25	DM (12)	DM (12)
EVIDENCE	24	Semantic (8)	lexical chain (7)
RESULT	21	Semantic (8)	lexical chain (7)
MOTIVATION	21	Semantic (8)	lexical chain (7)
PURPOSE	21	Syntactic (9)	infinitival clause (7)
CIRCUMSTANCE	20	DM (11)	DM (11)

Table 3: Distribution of Most Common Signals across Relations.

estingly, world knowledge such as the order of the presidents of the United States (e.g. that Bush served as the president of the United States before Obama) is also a indicative signal for SEQUENCE.

Another way of seeing these signals is to examine their associated tokens in texts, regardless of the signal types and subtypes. Table 4 lists some representative, generic/ambiguous (in boldface), and coincidental (in italics) tokens that correspond to the relations they signal. Each item is delimited by a comma; the & symbol between tokens in one item means that this signal consists of a word pair in respective spans. The number in the parentheses is the count of that item attested in this project; if no number is indicated, then that token span only occurs once. The selection of these single-occurrence items is random in order to better reflect the relevance in contexts. For instance, lexical items like *Professor Eastman* in JOINT, *NASA* in ELABORATION, *Bob McDonnell* in BACKGROUND, and *NATO* in RESTATEMENT appear to be coincidental because they are the topics or subjects being discussed in the articles. These results are parallel to the findings in Zeldes (2018a, p.180), which employed a frequency-based approach to show the most distinctive lexemes for some relations in GUM.

4.2 Distribution of Signals across Genres

Table 6 shows the distribution of the signaled relations in different genres. Specifically, the number preceding the vertical line is the number of signals indicating the relation and the percentage fol-

relations	examples of anchored tokens
JOINT	; (16), **and** (15), also (10), *Professor Eastman* (3), he (3), they (2)
ELABORATION	Image (6), based on (3), – (3), *NASA* (3), *IE6* (3), More specifically (2), Additionally (2), also (2), they (2), it (2), *Professor Chomsky* (2)
PREPARATION	: (6), How to (2), Know (2), Steps (2), Getting (2)
BACKGROUND	Therefore, Indeed, build on, previous, *Bob McDonnell*, Looking back
CONTRAST	**but** (9)/**But** (4), **or** (2), Plastic-type twist ties & paper-type twist ties, in 2009 & today, deteriorate & hold up, however, bad & nice, yet
RESTATEMENT	They & they (2), *NATO* (2), In other words, realistic & real, **and**, rehashed & retell, it means that, *Microsoft & Microsoft*
CONCESSION	**but** (10), However (3), The problem is (2), though (2), at least, While, It is (also) possible that, however, best & okay, Albeit, despite, if, still
JUSTIFY	because (2), an affront & disappointed deeply, excitement, share, *the straps*, The logic is that, any reason, **so**, **since**, confirm, inspire
EVALUATION	very serious, nationally representative, a frightening idea, a true win, an important addition, issue, This study & It, misguided, pain
SOLUTIONHOOD	Well (2), arrogant, :, **So**, why, **and**, *Darfur*, How, I think, Determine
CONDITION	If (12)/if(10), even if, unless, depends on, –, once, when, until
ANTITHESIS	**but** (5)/**But**, instead of (2), In fact, counteract, won't, rather than, **Or**, not, *the Arabs*, however, better & worst
SEQUENCE	**and** (3), First & Second, examined & assessed, later, Bush & Obama, initial, *digital humanities*, A year later, stop & update
CAUSE	because (3), suggests, due to, compensate for, **as**, **since/Since**, *arrogant people*, in turn, given, brain damage, as such
EVIDENCE	( ) (2), see (2), According to, because, **as**, –, **and**, *Arabs & Turkey*, Because of, *discrimination*, biases, The report states that, Thus
RESULT	**so** (3), **and** (2), meaning (2), so that, capturing, thus, putting, *the $\chi2$ statistic*, make
MOTIVATION	will (2), easier, *the pockets*, All it takes is, **so**, last longer
PURPOSE	to (6)/To, in order to (3)/In order to (2), **so** (2), enable, The aim
CIRCUMSTANCE	when (4)/When (2), On March 13, Whether, **As/as**, With, in his MIT office, the bigger & the harsher

Table 4: Examples of Anchored Tokens across Relations.

	examples of anchored tokens in different genres
academic	*discrimination* (16), **;** (11), and (8), **:** (5), **to** (5), but (5), also (5), though (3), **hypothesized** (3), **based on** (3), **First** & **Second** (3), however (3), because (2), More specifically (2), in/In order to (2), as (2), **( )** (2), **see** (2), when (2), **posited**, **expected**, capturing, **Albeit**
how-to guides	but (10), If (9)/if(7), **;** (5), and (4), also (4), *arrogant people* (9), **How** (7), **:** (3), so (3), **–** (3), But (3), **Know** (3), **Steps** (2), Move, Challenge, **Warnings**, In other words, Empty, Fasten, **Tips**, Wash
news	*IE6* (9), *NASA* (5), and (4), but (4)/But (2), **Image** (4), **market** (4) However (2), **the major source**, the Udvar-Hazy Center, in 2009
interviews	*Sarvis* (14), **What** (12), **Why** (11), and (8), *Noam Chomsky* (8), but (5), **Wikinews** (4), because (3), **interview** (2), **–** (2), **Well** (2), So (2), Which

Table 5: Examples of Anchored Tokens across Genres.

relations	academic	how-to guides	news	interview
JOINT	65 | 23.13%	76 | 18.67%	65 | 25.39%	54 | 16.77%
ELABORATION	61 | 21.71%	79 | 19.41%	53 | 20.70%	50 | 15.53%
PREPARATION	25 | 8.9%	55 | 13.51	15 | 5.86%	34 | 10.56%
BACKGROUND	33 | 11.74%	24 | 5.9%	28 | 10.94%	27 | 8.39%
CONTRAST	17 | 6.05%	21 | 5.16%	19 | 7.42%	11 | 3.42%
RESTATEMENT	N/A	20 | 4.91%	11 | 4.3%	29 | 9.01%
CONCESSION	17 | 6.05%	13 | 3.19%	10 | 3.91%	9 | 2.8%
JUSTIFY	1 | 0.36%	11 | 2.7%	15 | 5.86%	22 | 6.83%
EVALUATION	10 | 3.56%	12 | 2.95%	7 | 2.73%	13 | 4.04%
SOLUTIONHOOD	2 | 0.71%	8 | 1.97%	N/A	24 | 7.45%
CONDITION	N/A	25 | 6.14%	3 | 1.17%	3 | 0.93%
ANTITHESIS	3 | 1.07%	10 | 2.46%	1 | 0.39%	17 | 5.28%
SEQUENCE	12 | 4.27%	4 | 0.98	5 | 1.95%	5 | 1.55%
CAUSE	6 | 2.14%	12 | 2.95	6 | 2.34%	1 | 0.31%
EVIDENCE	10 | 3.56%	N/A	5 | 1.95%	9 | 2.8%
RESULT	3 | 1.07%	6 | 1.47%	6 | 2.34%	6 | 1.86%
MOTIVATION	N/A	21 | 5.16%	N/A	N/A
PURPOSE	14 | 4.98%	5 | 1.23%	N/A	2 | 0.62%
CIRCUMSTANCE	2 | 0.36%	5 | 1.23%	7 | 2.73%	6 | 1.86%
Total	281 | 100%	407 | 100%	256 | 100%	322 | 100%

Table 6: Distribution of Signaled Relations across Genres.

lowing the vertical line is the corresponding proportional frequency. The label N/A suggests that no such relation is present in the sample from that genre.

As can be seen from Table 6, how-to guides involve the most signals (i.e. 407 instances), followed by interviews, academic articles, and news. It is surprising to see that news articles selected from the GUM corpus are not as frequently signaled as they are in RST-SC, which could be attributed to two reasons. Firstly, the source data is different. The news articles from GUM are from Wikinews while the documents from RST-SC are Wall Street Journal articles. Secondly, RST-DT has finer-grained relations (i.e. 78 relations as opposed to the 20 relations used in GUM) and segmentation guidelines, thereby having more chances for signaled relations. Moreover, it is clear that JOINT and ELABORATION are the most frequently signaled relations in all four genres across the board, followed by PREPARATION in how-to guides and interviews or BACKGROUND in academic articles and news, which is expected as these four relations all show high-level representations of discourse that involve more texts with more potential signals.

Table 5 lists some signal tokens that are indicative of genre (in boldface) as well as generic and coincidental ones (in italics). The selection of these items follows the same criteria used in Section 4.1. Even though DMs *and* and *but* are present in all four genres, no associations can be established between these DMs and the genres they appear in. Moreover, as can be seen from Table 5, graphical features such as semicolons,

colons, dashes, and parentheses play an important role in relation signaling. Although these punctuation marks do not seem to be indicative of any genres, academic articles tend to use them more as opposed to other genres. Although some words or phrases are highly frequent, such as *discrimination* in academic articles, *arrogant people* in how-to guides, *IE6* in news, and *Sarvis* in interviews, they just seem to be coincidental as they happen to be the subjects or topics being discussed in the articles.

Academic writing is typically formal, making the annotation more straightforward. The results from this dataset suggest that academic articles contain signals with diverse categories. As shown in Table 5, in addition to the typical DMs and some graphical features mentioned above, there are several lexical items that are very strong signals indicating the genre. For instance, the verb *hypothesized* and its synonym *posited* are indicative in that researchers and scholars tend to use them in their research papers to present their hypotheses. The phrase *based on* is frequently used to elaborate on the subject matter. Furthermore, Table 5 also demonstrates that academic articles tend to use ordinal numbers such as *First* and *Second* to structure the text. Last but not least, the word *Albeit* indicating the relation CONCESSION seems to be an indicative signal of academic writing due to the register it is associated with.

How-to Guides are the most signaled genre in this dataset. This is due to the fact that instructional texts are highly organized, and the cue phrases are usually obvious to identify. As shown in Table 5, there are several indicative signal tokens such as the *wh*-word *How*, an essential element in instructional texts. Words like *Steps*, *Tips*, and *Warnings* are strongly associated with the genre due to its communicative needs. Another distinct feature of how-to guides is the use of imperative clauses, which correspond to verbs whose first letter is capitalized (e.g. *Know, Empty, Fasten, Wash*), as instructional texts are about giving instructions on accomplishing certain tasks and imperative clauses are good at conveying such information in a straightforward way.

News articles, like academic writing, are typically organized and structured. As briefly mentioned at the beginning of this section, news articles selected in this project are not as highly signaled as the news articles in RST-SC. In addition

to the use of different source data, another reason is that RST-DT employs a finer-grained relation inventory and segmentation guidelines; as a result, certain information is lost. For instance, the relation ATTRIBUTION is signaled 3,061 times out of 3070 occurrences (99.71%) in RST-SC, corresponding to the type *syntactic* and its subtype *reported speech*, which does not occur in this dataset. However, we do have some indicative signal tokens such as *market* and *the major source*.

Interviews are the most difficult genre to annotate in this project for two main reasons. Firstly, it is (partly) spoken language; as a result, they are not as organized as news or academic articles and harder to follow. Secondly, the layout of an interview is fundamentally different from the previous three written genres. For instance, the relation SOLUTIONHOOD seems specific to interviews, and most of the signal instances remain unanchored (i.e. 11 instances), which is likely due to the fact that the question mark is ignored in the current annotation scheme. As can be seen from Table 5, there are many *wh*-words such as *What* and *Why*. These can be used towards identifying interviews in that they formulate the question-answer scheme. Moreover, interviewers and interviewees are also important constituents of an interview, which explains the high frequencies of the two interviewees *Sarvis* and *Noam Chomsky* and the interviewer *Wikinews*. Another unique feature shown by the signals in this dataset is the use of spoken expressions such as *Well* and *So* when talking, which rarely appear in written texts.

5 Discussion

5.1 Annotation Scheme

For syntactic signals, one of the questions worth exploring is which of these are actually attributable to sequences of tokens, and which are not. For example, sequences of auxiliaries or constructions like imperative clauses might be identifiable, but more implicit and variable syntactic constructions are not such as ellipsis.

In addition, one of the objectives of the current project is to provide human annotated data in order to see how the results produced by machine learning techniques compare to humans' judgments. In particular, we are interested in whether or not contemporary neural models have a chance to identify the constructions that humans use to recognize discourse relations in text based on individual
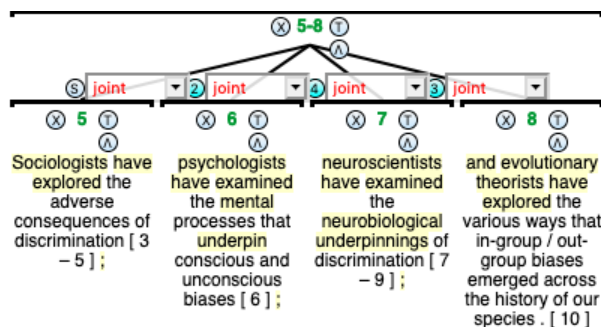

Figure 3: A Visualization of a Multinuclear Relation.

sequences of word embeddings, a language modeling technique that converts words into vectors of real numbers that are used as the input representation to a neural network model based on the idea that words that appear in similar environments should be represented as close in vector space.

Another dilemma that generally came up during the discussion about signal anchoring was whether or not to mark the first constituent of a multinuclear relation. In Figure 3, four juxtaposed segments are linked together by the JOINT relation, with associated signal tokens being highlighted. The first instance of JOINT is left unsignaled/unmarked while the other instances of JOINT are signaled. The rationale is that when presented with a parallelism, the reader only notices it from the second instance. As a result, signals are first looked for between the first two spans, and then between the second and the third. If there is no signal between the second and the third spans, then try to find signals in the first and the third spans. Because this is a multinuclear relation, transitivity does exist between spans. Moreover, the current approach is also supported by the fact that a multinuclear relation is often found in the structure like X, Y *and* Z, in which the discourse marker *and* is between the last two spans, and thus this *and* is only annotated for the relation between the last two spans but not between the first two spans. However, the problem with this approach is that the original source for the parallelism cannot be located.

5.2 Distribution of Discourse Signals

So far we have examined the distributions of signals across relations (Section 4.1) and genres (Section 4.2) respectively. Generally speaking, DMs are not only ambiguous but also inadequate as dis-

course signals; most signal tokens are open-class lexical items. More specifically, both perspectives have revealed the fact that some signals are highly indicative while others are generic or ambiguous. Thus, in order to obtain more valid discourse signals and parse discourse relations effectively, we need to develop models that take signals' surrounding contexts into account to disambiguate these signals.

Based on the results found in this dataset regarding the indicative signals, they can be broadly categorized into three groups: *register-related*, *communicative-need related*, and *semantics-related*. The first two are used to address genre specifications whereas the last one is used to address relation classification. Words like *Albeit* are more likely to appear in academic papers than other genres due to the register they are associated with; words like *Steps, Tips,* and *Warnings* are more likely to appear in instructional texts due to the communication effect they intend to achieve. Semantics-related signals play a crucial role in classifying relations as the semantic associations between tokens are less ambiguous cues, thereby supplementing the inadequacy of DMs.

5.3 Validity of Discourse Signals

It is also worth pointing out that some tokens are frequent signals in several relations, which makes their use very ambiguous. For instance, the coordinating conjunction *and* appears in JOINT, RE-STATEMENT, SEQUENCE, and RESULT in this dataset. Similarly, the subordinating conjunctions *since* and *because* serve as signals of JUSTIFY, CAUSE, and EVIDENCE in these 12 documents. These ambiguities would pose difficulties to the validity of discourse signals. As pointed out by Zeldes (2018a), a word like *and* is extremely ambiguous overall, since it appears very frequently in general, and is attested in all discourse functions. However, it is noted that some 'and's are more useful as signals than others: adnominal 'and' (example (5)) is usually less interesting than intersentential 'and' (example (6)) and sentence initial 'and' (example (7)).

(5) The owners, [William ***and*** Margie Hammack], are luckier than any others.[3] – ELABORATION-ADDITIONAL

[3]This example is chosen from the RST-DT corpus (Carlson et al., 2003) for illustration due to the apposition. Note that the relation inventory also differs.

(6) [Germany alone had virtually destroyed Russia, twice,]$_{n1}$ [***and*** Germany backed by a hostile military alliance, centered in the most phenomenal military power in history, that's a real threat.]$_{n2}$ – JOINT [interview_chomsky]

(7) [It arrests us.]$_N$ [***And*** then you say you won't commit a mistake, so you'll commit new mistakes. It doesn't matter.]$_S$ – ANTITHESIS [interview_peres]

Hence, it would be beneficial to develop computational models that score and rank signal words not just based on how proportionally often they occur with a relation, but also on how (un)ambiguous they are in contexts. In other words, if there are clues in the environment that can tell us to safely exclude some occurrences of a word, then those instances shouldn't be taken into consideration in measuring its 'signalyness'.

6 Conclusion

The current study anchors discourse signals across several genres by adapting the hierarchical taxonomy of signals used in RST-SC. In this study, 12 documents with 11,145 tokens across four different genres selected from the GUM corpus are annotated for discourse signals. The taxonomy of signals used in this project is based on the one in RST-SC with additional types and subtypes proposed to better represent different genres. The results have shown that different relations and genres have their indicative signals in addition to generic ones, and the indicative signals can be characterized into three categories: register-related, communicative-need related, and semantics-related.

The current study is limited to the rst annotation layer in GUM; it is worth investigating the linguistic representation of these signals through other layers of annotation in GUM such as coreference and bridging, which could be very useful resources constructing theoretical models of discourse. In addition, the current project provides a qualitative analysis on the validity of discourse signals by looking at the annotated signal tokens across relations and genres respectively, which provides insights into the disambiguation of generic signals and paves the way for designing a more informative mechanism to quantitatively measure the validity of discourse signals.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.

Debopam Das and Maite Taboada. 2017. Signalling of Coherence Relations in Discourse, Beyond Discourse Markers. *Discourse Processes*, pages 1–29.

Debopam Das and Maite Taboada. 2018. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Language Resources and Evaluation*, 52(1):149–184.

Luke Gessler, Yang Liu, and Amir Zeldes. 2019. A Discourse Signal Annotation System for RST Trees. In *Proceedings of 7th Workshop on Discourse Relation Parsing and Treebanking (DISRPT) at NAACL-HLT*, Minneapolis, MN. (To Appear).

Yang Liu and Amir Zeldes. 2019. Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Proceedings of the Society for Computation in Linguistics*, 2(1):314–317.

William C Mann and Sandra 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.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakesh, Morocco.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1023–1031. Association for Computational Linguistics.

Maite Taboada and Debopam Das. 2013. Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. *D&D*, 4(2):249–281.

Maite Taboada and Julia Lavid. 2003. Rhetorical and Thematic Patterns in Scheduling Dialogues: A Generic Characterization. *Functions of Language*, 10(2):147–178.

Maite Taboada and William C Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Amir Zeldes. 2016. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, pages 1–5, San Diego, CA.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2018a. *Multilayer Corpus Studies*. Routledge Advances in Corpus Linguistics 22. Routledge, London.

Amir Zeldes. 2018b. A Neural Approach to Discourse Relation Signaling. *Georgetown University Round Table (GURT) 2018: Approaches to Discourse*.