

# Unpacking Ambiguity: The Interaction of Polysemous Discourse Markers and Non-DM Signals

Jingni Wu

Georgetown University  
jw2175@georgetown.edu

Amir Zeldes

Georgetown University  
az364@georgetown.edu

## Abstract

Discourse markers (DMs) like ‘but’ or ‘then’ are crucial for creating coherence in discourse, yet they are often replaced by or co-occur with non-DMs (‘in the morning’ can mean the same as ‘then’), and both can be ambiguous (‘since’ can refer to time or cause). The interaction mechanism between such signals remains unclear but pivotal for their disambiguation. In this paper we investigate the relationship between DM polysemy and co-occurrence of non-DM signals in English, as well as the influence of genre on these patterns. Using the framework of eRST, we propose a graded definition of DM polysemy, and conduct correlation and regression analyses to examine whether polysemous DMs are accompanied by more numerous and diverse non-DM signals. Our findings reveal that while polysemous DMs do co-occur with more diverse non-DMs, the total number of co-occurring signals does not necessarily increase. Moreover, genre plays a significant role in shaping DM-signal interactions.

## 1 Introduction

Identifying and understanding discourse relations is fundamental to discourse comprehension. Discourse markers (DM) such as ‘and’, ‘because’, and ‘however’ have been widely recognized as the most typical indicator of coherence relations and are also referred to as discourse connectives or cue phrases (Forbes-Riley et al., 2006). Early research focused on DMs as the sole device indicating relations, and their presence is often used to distinguish explicit and implicit relations (Webber and Joshi, 1998; Robaldo et al., 2008). However, more recent studies have shown that DMs account for only a small fraction of discourse relations, which can be signaled by *reference* (e.g. anaphora to indicate ELABORATION<sup>1</sup>), *semantic* (antonymy to indicate CON-

TRAST), *lexical* (‘the next day’ can indicate temporal SEQUENCE like the DM ‘then’), *morphological* (past followed by present tense can also indicate SEQUENCE) and *graphical* cues (e.g. a question mark signaling a QUESTION relation).<sup>2</sup> Such non-DM signals can be crucial for disambiguating otherwise ambiguous DMs, such as ‘since’, which can signal both CAUSE and temporal CIRCUMSTANCE relations. Taken together, DMs and such similar non-DM devices are referred to collectively as discourse relation *signals* (Das and Taboada, 2018a,b; Zeldes et al., 2025).

Despite extensive research on DMs and other signals individually, far less attention has been given to their interaction. Prior studies have examined the distribution of DM-signal co-occurrence and explored potential motivations from corpus-based (Das and Taboada, 2019; Crible, 2020) and experimental perspectives (Crible and Demberg, 2020; Grisot and Blochowiak, 2017). These studies have revealed that DM-signal co-occurrence is influenced by cognitive constraints and information density, and that several factors, such as the ambiguity of DMs (Crible, 2020), the semantics of discourse relations (Das and Taboada, 2019; Crible and Demberg, 2020), and genres (Crible, 2020), affect the likelihood of co-occurrence. However, the specific mechanisms governing DM-signal interactions remain unclear. In particular, little is known about which conditions favor such co-occurrences, how different signals contribute to disambiguation and the resulting effect, what happens when conflicting signals appear, and how these patterns vary across discourse relations and genres.

While previous studies have confirmed that polysemous DMs co-occur with additional signals, there has been little systematic analysis of how different types and combinations of non-DM sig-

<sup>1</sup>Here and below we will assume discourse relation labels commonly used in Rhetorical Structure Theory (Mann and Thompson, 1988).

<sup>2</sup>For a full list of the non-DM signal types used in this study, see Appendix B.

nals help resolve ambiguity. This study seeks to bridge this gap by analyzing the distribution, number, type, and co-occurrence patterns of signals with polysemous DMs across genres. We focus on the following research questions:

1. Are polysemous DMs accompanied by more numerous or more diverse non-DMs?
2. What are the typical combination strategies for DM and non-DM signals?
3. Are strategies and distributions general, or are they genre-specific?

Because of their lower information content, we hypothesize that polysemous DMs will exhibit a stronger connection with non-DM prevalence. We also anticipate that different genres will exhibit distinct preferences for specific types of signals for polysemous DMs when resolving DM ambiguity, in part because they involve different prior likelihoods of certain relations. We therefore expect the relationship between DM polysemy and the number and diversity of co-occurring non-DMs to vary by genre.

## 2 Related Work

Previous studies have demonstrated that discourse relations are frequently signaled not just by DMs, with over 80% of signaled relations exhibiting some other textual cues, both with and without the presence of accompanying DMs (Taboada and Das, 2013; Das and Taboada, 2018a,b). Moreover, it has been found in many cases that multiple signals indicate discourse relations simultaneously (Das and Taboada, 2018b; Webber et al., 2019). Among these, the combined use of DMs and non-DM signals is particularly common and serves to signal a wide variety of relations (Das and Taboada, 2019). For instance, in the following example from the GUM corpus (Zeldes, 2017), ‘while’ functions as a typical DM for the CONCESSION relation, which is further reinforced by a lexical chain connecting existing ‘studies of the psychology of art’ with ‘no work’, creating a contrast between previous work that exists and a gap in academic literature:

- (1) [**While** studies of the psychology of art have focused on ... no work has been ...]  
[Relation: ADVERSATIVE-CONCESSION; DM: ‘While’; Signal: semantic (lexical chain)] (File: *GUM\_academic\_art*)

Although this pattern is very common in academic writing, little attention has been paid to the ways in which ambiguous DMs such as ‘while’ (which can also mean ‘during a time that...’) resolve to a unique interpretation thanks to co-occurring signals in this manner, and the joint use of DMs and signals remains a complex question.

Non-DM signals can 1) overlap with DMs in meaning, potentially leading to redundancy, 2) co-occur with DMs but function independently (potentially signaling multiple distinct relations), and 3) may complement DMs in specific types of relations and environments (Hoek et al., 2018). Recent studies have begun to explore the underlying triggers of the ‘DM + other signals’ phenomenon. Das and Taboada (2019) suggested that such combinations may arise from the inherent ambiguity of certain DMs which can signal various relations. For example, the DM *and* can mark additive LIST and temporal SEQUENCE relations, among other options, as illustrated in the following examples from GUM:

- (2) [I came home last night **and** told you.] [Relation: JOINT-SEQUENCE] (File: *GUM\_conversation\_grounded*)
- (3) [... borders of our moral **and** ethical understanding.] [Relation: JOINT-LIST] (File: *GUM\_essay\_ghost*)

Building on this, researchers have introduced the concept of ‘marking strength’ or ‘signaling strength’ of DMs, which can be assessed by the number and frequency of discourse relations they can signal (Asr and Demberg, 2012). Zeldes and Liu (2020) proposed the ‘delta-softmax’ metric, which quantifies prediction accuracy degradation for a trained neural model when a word is removed to estimate its signaling strength for a relation, providing empirical validation of an intuitive graded ‘signality-ness’ phenomenon. For instance, ‘but’ could be significantly less ambiguous than ‘and’ as a DM, in that removing ‘but’ would make the relation much harder to predict than removing ‘and’.

This strength directly influences how DMs interact with non-DM signals: it has been suggested that DMs tend to co-occur more frequently with other signals when indicating a wide range of discourse relations (Das and Taboada, 2019). In such cases, non-DM signals can play a disambiguation role, helping to clarify the intended relation (Crible and Demberg, 2020). However, although patterns

might be typical of specific genres, for example if formal texts prefer stronger and less ambiguous DMs, the association between DMs and other signals has not been found to vary significantly across genres in previous work (Crible, 2020).

In addition, combinations of DMs and non-DM signals vary across relation types, but they are not necessarily driven by inherent semantics (Das and Taboada, 2019). That is to say, certain relations tend to prefer either DM-only or DM-plus-signal combinations, which is partly influenced by the inherent semantics of the discourse relations themselves (e.g., weakly connected sentences), but also appears to reflect an independent pragmatic strategy for ensuring clarity of the writer’s intention.

While prior research has qualitatively identified some factors influencing the co-occurrences of DM and non-DM signals, a systematic analysis of how specific non-DM signals interact with ambiguous DMs across relation types and genres remains underexplored. In particular, the co-occurrence patterns between ambiguous DMs and accompanying signals have not been quantitatively mapped. Using the largest sample of annotated discourse relation signals to date, this study addresses this gap by investigating 1) the types and frequencies of non-DM signals that co-occur with ambiguous DMs, 2) how these combinations vary across genres, and 3) whether certain signal combinations contribute to disambiguating the intended discourse relation.

### 3 Data

This study uses the Georgetown University Multilayer (GUM) Corpus which consists of 16 spoken and written, informal and formal style English text types (Zeldes, 2017) (see corpus details in Appendix A). The corpus originally contained RST annotations, which were recently extended based on Enhanced Rhetorical Structure Theory (eRST, Zeldes et al. 2025), adding annotated DMs and seven types of non-DM signals based on the taxonomy proposed by Das and Taboada (Das and Taboada, 2018a), as well as adding multiple concurrent and tree breaking relations edges to the initial RST trees. With over 250,000 tokens, this is currently the largest dataset annotated for DMs and non-DM signals of discourse relations.

### 4 Polysemy of DMs

The ambiguous nature of DMs arises from their one-to-many relationship with discourse relations.

DMs that can signal multiple relations, such as ‘and’, are often described as ‘weak signals’ (Asr and Demberg, 2012; Das and Taboada, 2019; Crible, 2020), as they do not map consistently to a single meaning, in contrast to unambiguous DMs such as ‘despite’, which always marks a CONCESSION.

Going beyond previous categorical approaches to such polysemy, we adopt a graded, quantifiable definition of DM polysemy by calculating the Shannon Entropy (Shannon, 1951) of DM meanings, which measures how evenly a DM is distributed across multiple discourse relations. A high entropy score indicates that a DM appears equally in multiple relations, while a lower score means that a DM is used in only one or very few types of discourse relations, or with a strong predominant sense. We expect a high entropy score here for the most polysemous DMs, for example, a high value for DMs like ‘and’ or even ‘but’, and the lowest value for DMs like ‘despite’.

Shannon Entropy is calculated by measuring the probability of the DM appearing in each discourse relation. The polysemy score is computed as follows:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad (1)$$

$x_i$  is the possible discourse relation signaled by a DM,  $n$  is the number of distinct relations signaled by the DM, and  $P(x_i)$  is the probability of the DM signaling the relation  $x_i$ .

## 5 DM-Signal Co-occurrences

### 5.1 General Distribution

Across 16 genres, 21,435 discourse relations are annotated in our data, of which 1,372 (6.4%) are indicated by both DMs and non-DM signals. This result aligns fairly closely with Das and Taboada (2018a)’s finding for Wall Street Journal news (7.55%). However as suspected, we observe substantial variation across genres (see Figure 1): *essay* (8.7%), *bio* (8.6%), and *whow* (8.5%) show a higher proportion of DM-signal co-occurrence, whereas *conversation* has the lowest proportion (3.9%).

Among the 1,372 instances of DM-signal co-occurrence, 96% are marked by DM + 1 signal or DM + 2 signals, while just 3% are marked by three to four signals. Only a handful of cases include more than five signals (see Table 1).

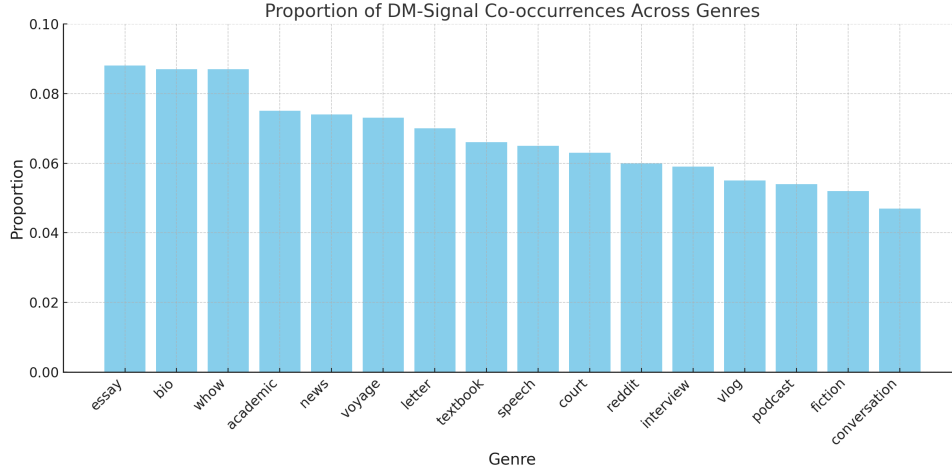


Figure 1: Proportion of DM-Signal co-occurrence across genres

	DM + 1 Signal	DM + 2 Signals	DM + 3 Signals	DM + 4 Signals	DM + 5 Signals	DM + 6 Signals	DM + 8 Signals
Total counts	1092	229	42	6	1	1	1
Proportion	79.55%	16.7%	3.1%	0.44%	0.07%	0.07%	0.07%

Table 1: Pattern of DM + signal combinations in co-occurrences

The most commonly used DM in co-occurrence with other signals across genres is the connective ‘and’ (36.6%), which is generally the most frequently used DM as well. Almost all genres in our corpus employ ‘and’ in DM-signal co-occurrences, except for *academic* where the conjunction ‘by’ is the most common DM favoring non-DM signal accompaniment, as in example (4), where the DM signaling the MEANS relation is accompanied by the lexical signal ‘using’:

- (4) **by** using a second order Rao and Scott (1981) ... correction

The top three most frequently used signal types in co-occurrences are ‘semantic’, ‘syntactic’, and ‘lexical’ across genres, though different genres favor different types of signals, in part due to the format of texts. In spoken genres such as *vlog*, *conversation*, and *court*, there is a large amount of ‘reference’ signals used along with DMs to indicate relations such as ELABORATION (see Figure 2). Trivially, ‘graphical’ signals such as quotation marks to signal ATTRIBUTION cannot occur in spoken language and are restricted to written data.

## 5.2 Polysemous DMs and Signal Patterns

The DM ‘so’ has the highest polysemy score across all genres in our dataset, while the DM ‘for’ ex-

hibits the most diverse range of accompanying signals (see Table 2). Here, ‘diversity’<sup>3</sup> refers to the number of distinct non-DM signal types that co-occur with a given DM, including individual signal types (e.g. ‘semantic’) and combinations of multiple types (e.g. ‘semantic + lexical’).

DM	non-DM signal diversity
<i>for</i>	29.50
<i>and</i>	26.64
<i>if</i>	25.00
<i>by</i>	20.80
<i>when</i>	19.00

Table 2: Top 5 DMs with the highest signal diversity

This raises the question of whether more polysemous DMs tend to co-occur with a greater number of non-DM signals and exhibit more diverse signal patterns, on account of the less consistent mapping of their form to a specific meaning. To answer these questions, we employed fitted regression models to examine the relationship between DM polysemy (independent variable) and two depen-

<sup>3</sup>Since DM frequency varies across genres, we normalized diversity by dividing the number of unique co-occurring signal types by the square root of total DM occurrences. This accounts for diminishing returns and prevents frequent DMs from being unfairly penalized.



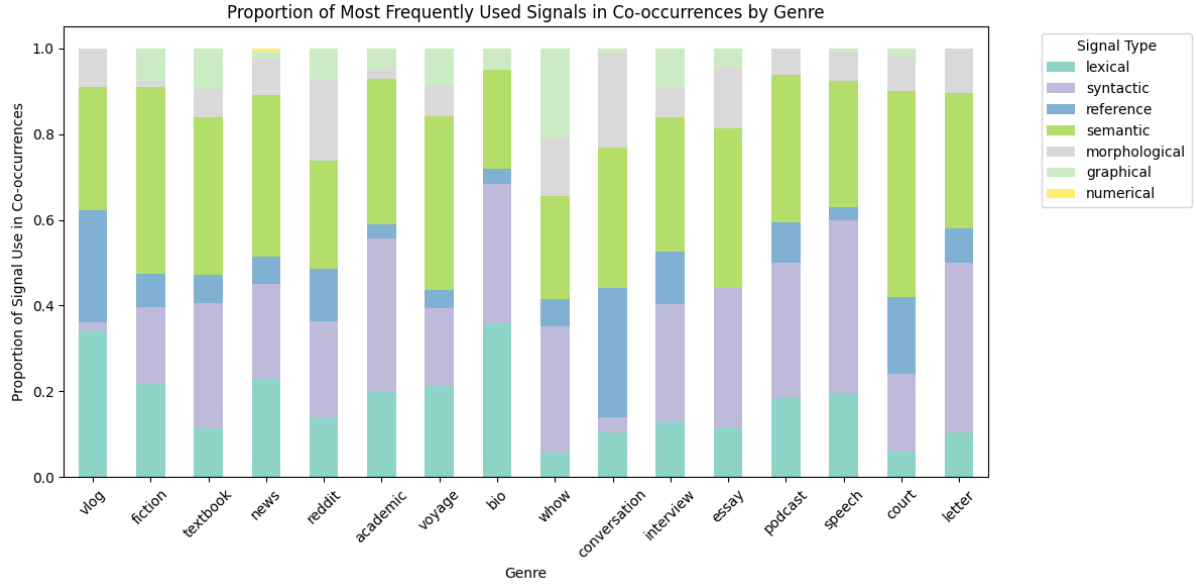


Figure 2: Proportion of most frequently co-occurring signals by genre

dent variables: (1) the total number of co-occurring non-DM signals and (2) the diversity of signal types associated with each DM.

Our results, based on both Pearson correlation and regression analyses (see details in Appendix C), suggest that polysemous DMs are more strongly associated with the diversity, rather than the quantity, of accompanying non-DM signals. While we observe a weak but statistically significant correlation between entropy and the total number of co-occurring signals ( $r = 0.248, p < 0.05$ ), this association does not hold in a multiple regression model where both entropy score and total co-occurring signals are included as predictors of normalized signal diversity. In contrast, entropy remains a significant predictor of normalized diversity, even after controlling for signal quantity ( $p < 0.001$ ). This supports the view that more polysemous DMs require more diverse signal patterns rather than just more signals to clarify their discourse functions.

However, the overall explanatory power of entropy score alone is modest (adjusted  $R^2 = 0.071$ ), suggesting that other factors may influence the relationship between DM polysemy and signal diversity. To further explore this, we considered genre as a variable. The regression model (see details in Appendix C) that includes genre and its interaction with entropy score significantly improved model fit ( $p < 0.000001$ , adjusted  $R^2 = 0.090$ ), suggesting that the effect of DM polysemy on co-occurring signal patterns varies across genres. Notably, genres such as *vlogs* exhibited a significantly stronger

positive relationship between DM polysemy and signal diversity, while others like *letter* showed a weaker or even negative trend. This variation highlights that the need for signal diversity in disambiguating polysemous DMs is not uniform, but shaped by genre-specific discourse norms. These genre-specific effects raise the question of what kinds of non-DM signal patterns are employed in each genre, which we address in the next section.

Looking at patterns rather than counts of signals in more detail, certain signal types consistently co-occur with highly polysemous DMs, suggesting that these signals play a crucial role in disambiguating them. For example, ‘lexical’ and ‘syntactic’ signals frequently appear across multiple cases and are more likely to be combined with other signal types, reinforcing their role in guiding interpretation (see Table 3).

In summary, our hypothesis is partially supported: polysemous DMs are more likely to exhibit diverse combinations of non-DM signal, possibly due to their less stable mapping of form to meaning, but they do not consistently co-occur with a greater number of signals. Given prior evidence that signal co-occurrences vary in quantity across genres (Figure 1), we now turn to investigate the impact of genre variation, and examine the hypotheses within individual genres in the following section.

### 5.3 Signal Combinations and Genre Effects

According to the entropy scores, the most polysemous DMs within each genre are presented in

DM	Top 3 co-occurring Types	Top 3 most frequent combinations
<i>so</i>	morphological, lexical, syntactic	(lexical + reference), (syntactic + reference + graphical)
<i>in</i>	syntactic, lexical	(syntactic + syntactic), (lexical + syntactic + syntactic)
<i>with</i>	semantic, graphical, syntactic	(reference + semantic), (syntactic + syntactic), (numerical + semantic + semantic)
<i>as</i>	syntactic, lexical, morphological	(lexical + semantic)
<i>and</i>	reference, lexical, semantic	(reference + graphical), (semantic + semantic), (lexical + syntactic)

Table 3: Top 5 Polysemous Discourse Markers and Co-occurring Signal Patterns

Table 4<sup>4</sup>. Notably, the most ambiguous DMs within each genre differs from those identified as globally most ambiguous. The DM ‘and’ is the most polysemous in six genres, and DM ‘so’ and ‘as’ are the second most ambiguous DMs in eight genres. By contrast, ‘also’ is the most polysemous DM in only one genre.

The non-DM signals that co-occur with polysemous DMs exhibit diverse combination patterns, which vary across genres. A single DM may be more likely to be paired with entirely different signals depending on the genre. For example, the DM ‘and’ is most frequently used with ‘lexical\_chain’ signals (a subtype of ‘semantic’) signals, see example (5)) in nearly all genres, except for *vlog*, *bio*, *whow*, *conversation*, and *podcast* (see Appendix Table 7). Here the lexical relation between the related items ‘information’ and ‘content’ forms a semantic signal next to ‘and’ to indicate that the two clauses are part of a list.

- (5) [The Penn State wiki was never proposed as a source of official information, **and** the university already hosts non-official content ...] [Relation: JOINT-LIST; DM: ‘and’; signal: semantic (lexical chain)](File: *GUM\_letter\_wiki*)

To further assess whether genre systematically affects the distribution of non-DM signals for polysemous DMs, we conducted Chi-Squared Goodness of Fit. For each genre, we compared the signal-type distribution to the global (genre-agnostic) distribution for the same set of DMs. After applying False Discovery Rate (FDR) correction, we found that all 16 genres show statistically significant deviations ( $p_{\text{corrected}} < 0.05$ ), confirming that genre has a strong effect on the signaling strategies used

to support polysemous discourse markers. Genres such as *vlog* and *conversation* exhibited the largest deviations, suggesting that signal use in these genres is especially distinct from the overall norm.

This variation can be attributed to the nature of spoken genres such as *vlogs* and *conversations*, which emphasize audience interaction and shared common ground. In these contexts, indicative words and personal references are more commonly used to enhance engagement and coherence. Similarly, other spoken genres tend to favor ‘reference’ signals, particularly ‘personal references’, using chains of pronouns to help the audience recall previously mentioned content. Semantic signals in the genre *podcast* show a particularly strong use of *meronymy*, using words in a part-whole relationship alongside the polysemous ‘and’ to indicate elaborations on complex information.

In addition, ‘and’ tends to use combined signals more frequently than other polysemous DMs, dovetailing with our initial hypothesis about non-DMs compensating for ambiguous DMs. Notably, in almost all genres where ‘and’ is the most polysemous DM, it co-occurs with multiple signals, except for the genre *essay*, where it primarily appears by itself or with a single signal type. Among all signal combinations, the most frequent combined signal set for ‘and’ is ‘reference’ + ‘semantic’, i.e. anaphora and lexical relations between words in the units joined by ‘and’. Interestingly, *letter* is the only genre where the most polysemous DM is ‘as’, yet it does not co-occur with any additional non-DM signals. Looking at its instances, nearly 65% are used to indicate MODE relations (manner/means), as opposed to only 32.2% in the rest of the corpus, suggesting that this usage may simply be more predictable as a default in *letters* – the most common sense in the remaining genres is indicating a temporal CIRCUMSTANCE, similarly to ‘when’.

Many DMs exhibit reduced polysemy within in-

<sup>4</sup>When comparing the polysemy across genres, we normalized the entropy score by dividing the raw entropy score by the maximum possible entropy for each DM in each genre.

Genre	DM	Raw entropy	Normalized entropy
Court	and	2.85	0.61
Reddit	so	2.59	0.60
Conversation	and	2.63	0.58
News	as	2.55	0.57
Fiction	so	2.35	0.53
Voyage	as	2.33	0.53
Interview	and	2.34	0.52
Vlog	and	2.27	0.52
Speech	so	2.20	0.51
Wikihow	so	2.25	0.50
Textbook	so	2.16	0.48
Podcast	and	2.15	0.48
Biography	also	1.90	0.44
Academic	as	1.92	0.44
Letter	as	1.84	0.42
Essay	and	1.64	0.40

Table 4: Entropy score of DMs per genre

dividual genres compared to their global scores, suggesting that their meaning is more specialized and thus less ambiguous in certain contexts. However, some DMs show substantial variation across genres, potentially requiring a greater variety or higher number of non-DM signals to aid interpretation in specific genres (see Figure 3).

To identify DMs whose polysemy varies the most across genres, we compared their normalized within-genre polysemy scores with their global polysemy scores. The top five discourse markers with the largest shifts are ‘so’, ‘in’, ‘with’, ‘given’, and ‘indeed’, which align with the overall polysemy ranking observed earlier. Highly polysemous DMs exhibit greater variance across genres, likely because their multiple meanings make them more adaptable to different discourse needs, which can be disambiguated either by non-DM signals, or simply by their use in a genre with strong priors on expected senses. In contrast to DMs with lower polysemy, which may serve more stable functions, highly polysemous DMs can shift more dramatically depending on genre-specific discourse structures, discourse relation compatibility, communicative conventions, and signaling strategies.

Among the genres, *academic*, *reddit*, and *court* seem to have larger variance, indicating that DMs used in these genres experience the most significant shifts in polysemy compared to their global usage. These genres may have DMs that behave very differently in terms of polysemy compared to their global usage. In contrast, DMs in *fiction*, *podcast*, and *letter* appear to behave similarly locally and globally.

In addition, we examined the relationship between the number of co-occurring non-DM signals, the diversity of those signals, and the DM polysemy within each genre. Our global analysis confirms that polysemous DMs tend to co-occur with more diverse signal patterns, however, since the frequency and variety of co-occurring signals differ across genres, we extended this investigation within individual genres to determine whether genre influences this phenomenon. The results indicate that spoken genres such as *court*, *podcast*, and *vlog*, DM polysemy strongly correlates with both the number and diversity of co-occurring non-DM signals, while other genres’ results almost align with our findings across genres, that the higher a DM’s polysemy score, the more diverse these signal combinations tend to be. This supports our hypothesis that specific genres, particularly spoken contexts and formal or unusual settings (e.g. courtroom transcripts or academic writing), adopt distinct non-DM signaling strategies which help in the disambiguation of polysemous DMs.

## 6 Discussion

This study investigates the relationship between DM polysemy, the number and diversity of co-occurring non-DM signals, and the role of genre in these interactions. Our findings partially support the hypothesis that polysemous DMs exhibit more diverse non-DM signal patterns but do not necessarily co-occur with a greater number of non-DM signals. Moreover, genre greatly shapes DM polysemy, with significant variations in DM entropy and signal usage.

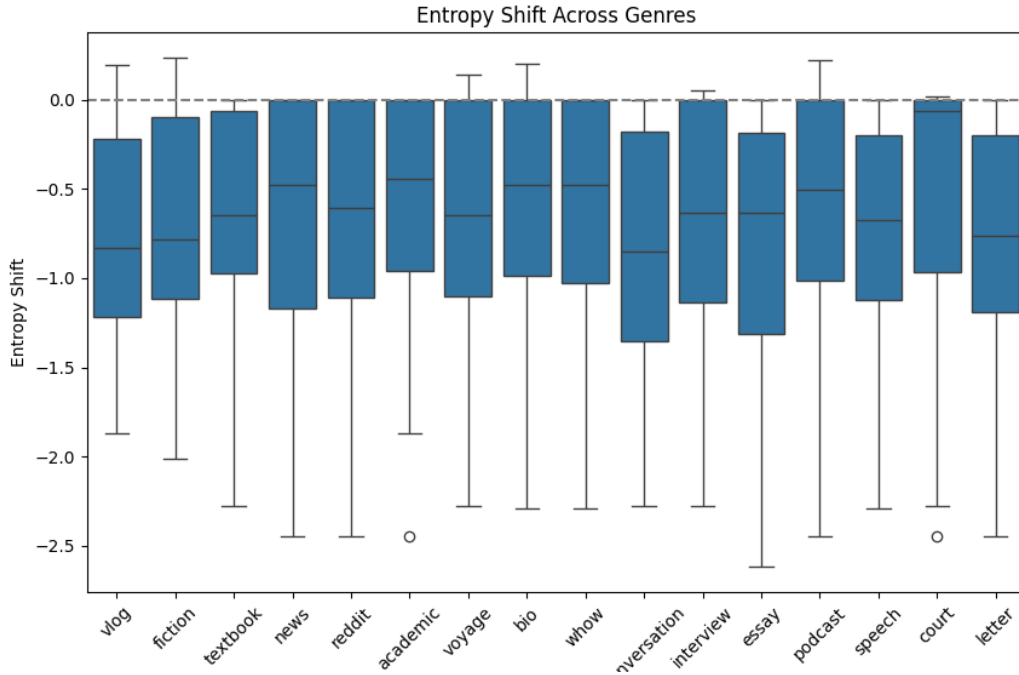


Figure 3: Entropy shift across genres

The regression analysis suggests that polysemous DMs are more likely to co-occur with more diverse non-DM signals. However, this relationship varies by genre, with spoken genres (e.g. court, podcast, vlog) showing a stronger dependence on non-DM signals to disambiguate polysemous DMs. This may reflect the unique cognitive and interactional constraints of speech. Spoken discourse requires real-time processing, and speakers often experience pressure to maintain fluency while managing limited cognitive resources for lexical retrieval (Clark, 2002). As a result, they may rely more on non-DM signals to support discourse coherence, even when these are partially redundant. Furthermore, DMs in spoken texts often serve interactive and procedural functions beyond marking semantic relations, such as managing turn-taking or structuring discourse (Clark and Tree, 2002). This functional difference may explain why speech genres exhibit stronger correlations between DM polysemy and signal use (numbers and types), as speakers may compensate for ambiguity and uncertainty in perception through additional diverse discourse cues.

On the other hand, some genres exhibit little to no significant correlation among DM polysemy, the number and the diversity of non-DM signals, suggesting that different discourse contexts may

impose different constraints on how DMs interact with non-DM signals. Additionally, we identified DMs whose polysemy scores are highly shifted across genres, such as, ‘so’, ‘in’, ‘with’, ‘given’, ‘indeed’, and ‘while’. This finding suggests that certain polysemous DMs are more sensitive to contextual variation, whereas others maintain stable meanings across different discourse settings.

This study does not fully account for the distribution of different discourse relations, which can further shape the observed patterns of polysemy and signal co-occurrence. Prior research has demonstrated that certain non-DMs are more commonly used to disambiguate DMs in specific relations, such as *contrast* and *consequence* (Crible and Demberg, 2020), and different relations may vary in their sensitivity to signals, with some relations being more reliant on co-occurring non-DM cues for disambiguation. Moreover, the compatibility between DMs and specific signals may play a greater role in guiding interpretation than sheer signal frequency. Future research could explore how different discourse relations condition the use of non-DM signals with polysemous DMs, providing a more fine-grained understanding of signal-based disambiguation mechanisms.



## References

- Fatemeh Torabi Asr and Vera Demberg. 2012. Measuring the strength of linguistic cues for discourse relations. *Proceedings of the Workshop on Advances in Discourse Analysis and Its Computational Aspects*, pages 33–42.
- Herbert H Clark. 2002. Speaking in time. *Speech Communication*, 36(1-2):5–13.
- Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.
- Ludivine Crible. 2020. Weak and strong discourse markers in speech, chat, and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes*, 57(9):793–807.
- Ludivine Crible and Vera Demberg. 2020. The role of non-connective discourse cues and their interaction with connectives. *Pragmatics & Cognition*, 27(2):313–338.
- Debopam Das and Maite Taboada. 2018a. Rst signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149–184.
- Debopam Das and Maite Taboada. 2018b. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770.
- Debopam Das and Maite Taboada. 2019. Multiple signals of coherence relations. *Discours*, (24).
- Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in d-ltag. *Journal of Semantics*, 23(1):55–106.
- Cristina Grisot and Joanna Blochowiak. 2017. [Temporal connectives and verbal tenses as processing instructions: Evidence from french](#). *Pragmatics & Cognition*, 24:404–440.
- Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2018. The linguistic marking of coherence relations: Interactions between connectives and segment-internal elements. *Pragmatics & Cognition*, 25(2):276–309.
- William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Alan Robaldo, Eleni Miltsakaki, Alan Lee, Rashmi Prasad, Nikhil Dinesh, Bonnie Webber, and Aravind Joshi. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. ELRA.
- Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64.
- Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *University of Pennsylvania*, 35:108.
- Bonnie Lynn Webber and Aravind K Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. *arXiv preprint cmp-lg/9806017*.
- Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.
- Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. [erst: A signaled graph theory of discourse relations and organization](#). *Computational Linguistics*, 51(1):23–72.
- Amir Zeldes and Yang Janet Liu. 2020. [A neural approach to discourse relation signal detection](#). *Dialogue and Discourse*, 11(2):74–99.

## A GUM Information

Text type	Source	Docs	Tokens
Academic writing	Various	18	17,169
Biographies	Wikipedia	20	18,213
CC Vlogs	YouTube	15	16,864
Conversations	UCSB Corpus	15	17,932
Courtroom transcripts	Various	9	11,148
Essays	Various	9	10,842
Fiction	Various	19	17,511
Forum	reddit	18	16,364
How-to guides	wikiHow	19	17,081
Interviews	Wikinews	19	18,196
Letters	Various	12	9,989
News stories	Wikinews	24	17,186
Podcasts	Various	10	11,986
Political speeches	Various	15	16,720
Textbooks	OpenStax	15	16,693
Travel guides	Wikivoyage	18	16,515
<b>Total GUM</b>		255	250,409

Table 5: Overview of GUM corpus by text type.

## B Non-DM signal types

In this paper we follow the taxonomy of non-DM signal types proposed by [Zeldes et al. \(2025\)](#), which distinguishes eight major classes with a total of 45 subtypes, illustrated in Table 6.

## C Regression Results

signal type	subtypes	example
dm	but, then, on the other hand...	[They wanted to] [ <b>but</b> couldn't] <adversative-contrast>
graphical	colon, dash, semicolon layout items in sequence parentheses, quotation marks question mark	[Let me tell you a story :] <organization-preparation> [Introduction] <organization-heading> 1. wash [2. cut] <joint-list> it rained [(and snowed a bit)] <elaboration-additional> [Did you?] <topic-question> No.
lexical	alternate expression indicative word/phrase	He agreed. [ <b>That is he</b> said yes] <restatement-repetition> They planned a party! [ <b>That's nice/Can't wait!</b> ] <evaluation-comment>
morphological	mood tense	<b>Go</b> with them [I think you should] <explanation-motivation> <b>I started</b> an hour ago, [now I'm resting] <joint-sequence>
numerical	same count	[ <b>Two</b> reasons.] <organization-preparation> <b>First...</b>
reference	comparative demonstrative / personal propositional	[I don't want <b>it</b> ] <adversative-antithesis> <b>I want another one.</b> They met <b>Kim</b> . [ <b>This person / she</b> was...] <elaboration-additional> <b>They met Kim</b> . [ <b>This encounter</b> was...] <elaboration-additional>
semantic	antonymy attribution source lexical chain meronymy negation repetition/synonymy	Beer is <b>cheap</b> , [wine is <b>expensive</b> ] <adversative-contrast> [ <b>Kim</b> said] <attribution-positive> <b>they</b> would it was <b>funny</b> [so they <b>laughed</b> ] <causal-result> <b>The house</b> was big, [ <b>the door</b> two meters tall] <elaboration-additional> <b>Kim</b> danced, [Yun <b>didn't</b> dance] <adversative-contrast> They met <b>Dr. Kim</b> . [ <b>Dr. Kim/The surgeon</b> was...] <elaboration-additional>
syntactic	infinitival/relative clause interrupted matrix clause modified head nominal modifier parallel syntactic construction past/present participial clause reported speech subject auxiliary inversion	a plan [ <b>to</b> win] <purpose-attribute> [I meant -] <organization-phatic> <b>I</b> mean, a <b>plan</b> [to win] <purpose-attribute> articles [ <b>explaining</b> chess] <elaboration-attribute> <b>it's all</b> tasty [ <b>it's all</b> pretty] <joint-list> <b>Kim</b> appeared [ <b>dressed in black</b> ] <elaboration-attribute> [Kim said] <attribution-positive> <b>that they would</b> I would have [ <b>had</b> I known] <contingency-condition>

Table 6: Signal types and subtypes, with examples highlighting in red the signal tokens which indicate the relation of the unit in square brackets.

Table 7: Signal Patterns for "and" by Genre

Genre	Top 3 "and" + 1 signal		Top 3 "and" + multiple signals	
	Signal Type	Signal Subtype	Signal Type	Signal Subtype
<b>vlog</b>	lexical semantic reference	indicative_word lexical_chain personal_reference	reference + semantic reference + reference reference + semantic	personal_reference + lexical_chain oral_reference + propositional_reference personal_reference + synonymy
<b>textbook</b>	semantic semantic graphical semantic	lexical_chain meronymy semicolon lexical_chain	graphical + graphical graphical + reference semantic + semantic + semantic + semantic semantic + semantic	items_in_sequence + semicolon parentheses + personal_reference lexical_chain + lexical_chain + lexical_chain + lexical_chain lexical_chain + meronymy
<b>reddit</b>	lexical reference semantic	indicative_word personal_reference lexical_chain	reference + reference + semantic reference + semantic + semantic reference + semantic	personal_reference + propositional_reference + synonymy personal_reference + lexical_chain + repetition personal_reference + meronymy
<b>academic</b>	lexical semantic semantic	indicative_word meronymy lexical_chain	lexical + lexical graphical + graphical + semantic semantic + semantic	indicative_word + indicative_word items_in_sequence + semicolon + meronymy lexical_chain + lexical_chain
<b>voyage</b>	lexical semantic lexical	meronymy indicative_word indicative_word	lexical + lexical semantic + semantic lexical + lexical	indicative_phrase + indicative_word lexical_chain + meronymy indicative_word + indicative_word
<b>bio</b>	lexical semantic lexical graphical	lexical_chain indicative_phrase items_in_sequence	lexical + lexical semantic + semantic semantic + semantic	indicative_phrase + indicative_word lexical_chain + lexical_chain lexical_chain + meronymy
<b>whow</b>	semantic reference reference	lexical_chain personal_reference personal_reference	graphical + semantic semantic + semantic reference + reference	items_in_sequence + lexical_chain lexical_chain + lexical_chain personal_reference + personal_reference
<b>conversation</b>	morphological semantic semantic	tense lexical_chain lexical_chain	reference + semantic reference + semantic semantic + semantic	personal_reference + personal_reference personal_reference + synonymy personal_reference + lexical_chain
<b>fiction</b>	semantic lexical semantic	meronymy indicative_word lexical_chain	graphical + lexical semantic + semantic semantic + semantic	lexical_chain + meronymy semicolon + indicative_word lexical_chain + lexical_chain
<b>news</b>	semantic lexical semantic	meronymy indicative_phrase lexical_chain	lexical + morphological semantic + semantic lexical + lexical + lexical	lexical_chain + meronymy indicative_word + tense lexical_chain + lexical_chain
<b>interview</b>	reference graphical	personal_reference semicolon	semantic + sementic + semantic semantic + semantic	indicative_word + indicative_word + indicative_word lexical_chain + lexical_chain + meronymy lexical_chain + synonymy
<b>essay</b>	semantic semantic lexical	lexical_chain meronymy alternate_expression	lexical + lexical + lexical	indicative_word + indicative_word + indicative_word
<b>podcast</b>	semantic reference lexical	meronymy personal_reference indicative_word	reference + reference + semantic semantic + semantic lexical + lexical	demonstrative_reference + personal_reference + meronymy lexical_chain + synonymy indicative_word + indicative_word
<b>speech</b>	lexical semantic syntactic	lexical_chain meronymy parallel_synatactic_construction	lexical + lexical reference + reference	indicative_word + indicative_word personal_reference + personal_reference
<b>court</b>	semantic reference semantic	lexical_chain personal_reference negation	reference + semantic reference + reference + semantic reference + semantic	demonstrative_reference + synonymy personal_reference + personal_reference + synonymy personal_reference + lexical_chain
<b>letter</b>	semantic reference semantic	lexical_chain personal_reference meronymy	reference + reference	personal_reference + personal_reference

Table 8: Pearson Correlation: Entropy Score and Total Co-occurred Signals

<b>Variable Pair</b>	<b>Correlation (r)</b>	<b>p-value</b>
Entropy Score - Total Co-occurred Signals	0.248	0.0137



Table 9: Model 1: Regression of Entropy Score on Normalized Signal Diversity

	<b>Coefficient</b>	<b>Std. Error</b>	<b>p-value</b>
<b>Intercept</b>	0.850	0.040	<0.001
<b>Entropy Score</b>	0.112	0.039	0.005
$R^2$		0.081	
Adjusted $R^2$		0.071	
F-statistic	8.44	$(p = 0.0046)$	

Table 10: Model 2: Regression of Entropy Score and Total Signals on Normalized Signal Diversity

	<b>Coefficient</b>	<b>Std. Error</b>	<b>p-value</b>
<b>Intercept</b>	0.851	0.040	<0.001
<b>Entropy Score</b>	0.123	0.040	0.003
<b>Total Co-occurred Signals</b>	-0.0005	0.0005	0.302
$R^2$		0.091	
Adjusted $R^2$		0.072	
F-statistic	4.76	$(p = 0.0107)$	

Table 11: Model 3: Regression of Within-Genre Entropy and Genre Interaction on Normalized Signal Diversity

<b>Coefficient</b>	<b>Coef.</b>	<b>Std. Error</b>	<b>p-value</b>
<b>Intercept</b>	0.866	0.034	<0.001
<b>Within-Genre Entropy</b>	0.055	0.033	0.098
Entropy $\times$ Genre[T.vlog]	0.132	0.050	0.009
Entropy $\times$ Genre[T.letter]	-0.119	0.061	0.053
Entropy $\times$ Genre[T.conversation]	0.075	0.047	0.116
Entropy $\times$ Genre[T.reddit]	0.073	0.047	0.123
Entropy $\times$ Genre[T.fiction]	0.077	0.055	0.162
Entropy $\times$ Genre[T.speech]	-0.066	0.049	0.178
$R^2$		0.121	
Adjusted $R^2$		0.090	
F-statistic	3.97	$(p < 0.000001)$	