Time as Rhetoric: Relating Temporal Order and Directed Narrative Intent in the Clinical Text Domain

Luke M. Breitfeller

CMU-LTI-26-001

February 2026

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carolyn P. Rosé, Carnegie Mellon University (Chair)
Lori Levin, Carnegie Mellon University
Teruko Mitamura, Carnegie Mellon University
Mark Dredze, Johns Hopkins University
Denis Newman-Griffis, University of Sheffield

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technology.

Acknowledgments

Completing a dissertation is a work of passion and struggle. There were days it seemed like the work would never be done, that it would never be meaningful, that it would never contribute anything to the field. I owe my perseverance to those around me–my mentors, my peers, and my loved ones. Without them, the work would never have been completed.

To my academic mentors: my advisor, Carolyn Rosé, who pushed me forward at every step and stumble, and helped me make the connections that would shape my academic journey. To my committee members, who saw the potential in the dissertation thesis and offered insight on how I could speak its message clearly. To Beth Rasch and the NIH EpiBio research team, who investigated important problems in clinical text with irreplaceable camaraderie. To my professors, who gave me a foundation from which to build. My research career has taken me places I never would have imagined at the start–but my mentors believed in me, even then. I would not be the researcher I am today without your guidance.

To my peers: classmates, fellow advising group members, and research collaborators. Many of you were in the same position I was when we met–still learning, still developing the skills to stand on our own as academics. And yet, I felt supported by every one of my peers in the LTI. This dissertation would not be possible if not for every project I had worked on before it; I build on the work of passionate and talented researchers before me, just as I hope those who come after will build on mine. Academia is collaboration, and I am grateful to have been able to participate in that process with you.

To my loved ones: my family, who have had more faith in me than I had in myself. To my friends far and near, the journey would be more difficult without the times of light and laughter you shared with me. To my partner, Michael, who has been by my side through every triumph and hardship, whose support has meant more than I have the words for. You have all met each new step in my dissertation process with excitement, and reminded me why I should feel pride in what I have achieved. Friends, family, and other loved ones–you are what makes this achievement today worthwhile.

To everyone who has helped me as I completed this dissertation, in whatever way: thank you. You are the heart of this work.

Contents

List of Figures

List of Tables

Abstract

Identifying the temporal arrangement of events within narrative text remains a significant challenge in natural language processing. Intuition would suggest that the simplest way to convey episodic memories is through chronological order, but human writers often deviate from fully-ordered timelines. Despite decades of work in the field, state-of-the-art models have significant gaps around the complex reasoning required to resolve these discrepancies, especially for known challenge cases like long-distance event-pairs. This dissertation theorizes that one gap may be explained by contextual (rather than textual) elements: the intentions of the human author at the time of text construction.

The work's overarching goal is to identify the relationship between the *purpose* of a text and the *arrangement of time* in its content. To that end, it develops a new corpus scoped to capture authorial intention and analyzes the corpus output to extract trends in temporal behavior. The dissertation findings do suggest a correlation between these two factors: chronological texts reflect a need by the text's author to be clear and informative, while efforts to invoke emotion or persuade an audience feature more frequent deviations from chronological time. Despite this, it is difficult to integrate narrative intent into temporal reasoning models. The qualitative observations do not, as yet, translate into quantitative improvements over SOTA, but they indicate possibilities for future work.

The dissertation proposes end-to-end methodology for examining this relationship between authorial intent and in-text temporal order. The methodology begins with a framework through which the factors that shape a text's relationship with chronology (both internal and external to the author) can be identified. It defines schemas for annotation of comprehensive temporal order even in low-resource studies, and designs a survey to compare temporal order in text against explicit markers of authorial intention. The survey data is collected and examined using qualitative and quantitative methods. Finally, experiments are run to measure the impact of directly looping intention features into predictive models for temporal order. It assembles and contributes annotation schemas, a corpus of narrative texts sorted by the intentions of the author (called IlNeS), and experimental settings which integrate the work into an existing SOTA temporal ordering model (known as MulCo, with the expanded model called TempR-MInt).

1 Introduction

Time is fundamental to textual comprehension; to understand how a sequence of events described in text occurred in real time, a reader must understand how that time is encoded in the text creation step.

The details of time and their impact on the final text are a critical focus of this dissertation work. My own graduate and undergraduate career has a focus on these nuances: specifically, much of my work has been shaped towards understanding how humans convey time within a constructed text document. The dissertation takes a specific interest in **temporality as an element of narrative**–though competing natural language processing models possess the ability to internalize complex patterns of text, these distinct motivations that exist within narrative text construction are an underused element of temporal analysis. The approach of the dissertation is informed by prior hands-on work in both annotation and extraction, where specialized models were scoped to capture time knowledge hidden in distinct *layers* of a textual narrative.

This dissertation sits at an intersection of often-disparate fields of study. Natural language processing, as a field, does benefit from understanding how human text is generated–generally, these insights are drawn from the related fields of linguistics (computational linguistics in particular), rhetoric, and discourse. While the dissertation draws insight from these, it chooses

to further incorporate *literary analysis*, *narrative theory* and *creative writing* (in domains of both fiction and non-fiction) as knowledgeable sources on the human elements of time. These fields provide unique insights about **human text generation as a motivated process**, and positions the dissertation as having an unusual viewpoint within the field. This work takes a view of text-based communication which centers the human element, and argues that this perspective complements rather than contradicts the statistical approaches more common in NLP. The dissertation's contributions act to bridge these two views of time within text–time as a tool of narrative, and time as a problem in computer science–and demonstrate how this processing of bridging can further advance both sides.

Temporal Deviation as a Tool of Narrative:

Temporal elements of text can roughly be divided into two categories: chronological and non-chronological. This dissertation studies the motivations behind non-chronological temporal elements, called **temporal deviations**. If the purpose of a text is to communicate the temporal ordering of events clearly, intuition suggests that texts should be arranged chronologically. But in art and literature, it is understood that targeted uses of temporal deviations often strengthen narrative. Non-chronological instances of time can enhance mood, convey theme, and make complex topics more comprehensible than if conveyed in chronological order; this tension between temporal types is one of the factors which makes effective narrative writing so challenging to humans. In the *The Art of Writing Nonfiction*, considered a foundational instructive text for writing creative nonfiction, Fontaine notes that "our training pushes us to chronological organization and the requirements of comprehension prohibit it" (Fontaine et al., 1987). The dissertation uses "narrative" as a focus to better understand the relationship between the function of text and time within it, with an eye towards improving current computational models.

Temporal Deviation in Computer Science:

"Temporal reasoning" covers a broad range of reading comprehension tasks within computational machine learning, but one basic building block of temporal reasoning is **order between events**. Models considered state-of-the-art for temporal ordering before LLMs still demonstrate gaps in their ability to perform temporal ordering and organize full timeline sequences. Current LLMs are flexible and capable of some temporal reasoning, but struggle with challenge cases their more-specialized predecessors succeed at. General improvements have been suggested to fill those gaps (e.g. multi-shot, chain-of-thought), with mixed results. This dissertation proposes a different approach, building on a narrative-driven understanding of time to improve state-of-the-art. A specific focus in **Temporal Event-pair Ordering**, also known as **Event Temporal Relation Extraction**, examines the utility of this approach for downstream temporal reasoning work. One aim of this work is to provide models which improve on a **TEO/ETRE** baseline. However, this dissertation also discusses flaws common to TEO/ETRE metrics, and illuminates that current SOTA models (despite scoring well on these metrics) perform poorly for temporal deviations. Moments of deviation within text represent a specific challenge case for temporal reasoning, and the dissertation seeks to refine the TEO/ETRE framework to better surface and predict for these deviations.

Audience and Contributions:

The contributions of this dissertation build from and are useful to multiple areas of research. It studies time within text, uses the approach of narrative analysis, and grounds itself with a case study in the clinical domain. We consider the audiences of this work to be:

- Linguists and computational linguists who are interested in evaluating the relationship

between *motivated intention* on the part of a text's author and the *temporal arrangement* of events within that text.

- Machine-learning engineers and NLP researchers developing models within the temporal reasoning and similar spheres. The methodology and results of the dissertation seek to fill gaps in existing SOTA.
- Theorists and researchers performing qualitative analyses of narrative. This dissertation provides novel insights based on the new dataset, but also scaffolds existing theories on narrative with more rigorous data. This subset of the dissertation audience also includes theorists specializing in **illness narrative**.
- Researchers examining the clinical domain. The survey collected by this dissertation simulates the experience of *patient-clinician communication* in certain cases and could provide insight into that element of the healthcare experience.

The dissertation contributes the following to research on time, narrative, and clinical text analysis.

1. **STAGE**, a temporal grammar extraction tool proposed to better suit the needs of complex temporal reasoning. This work can be used to extract features helpful to ML models across multiple temporal tasks.
2. An annotation framework which reduces the time and cognitive load required for high-quality TEO/ETRE annotations. This methodology offers utility for data generation across a variety of temporal reasoning tasks downstream to TEO/ETRE in ML and NLP.
3. **IlNeS**, a 106-document corpus of narratives generated by human participants, with annotations for events, time, and the authorial intent at time of document creation. This corpus is to be released in anonymized format for public use through GitHub[1]. This corpus provides direct training data for the TEO/ETRE task and accumulates information about narrative intention, time within text, the genre of illness narrative, and individual patient experiences with chronic or acute illness and disability. It represents significant potential for future study across all identified audiences of the dissertation as a whole.
4. Analysis using multiple metrics (qualitative and quantitative) of the relationship between authorial intent and time within the IlNeS text corpus. Insights from this analysis may be useful in narrative theory, computational linguistics, and ML/NLP.
5. An analysis of specific trends within IlNeS testimonies identified as having 'clinical' intent. This analysis may be helpful to researchers studying patient-clinician communication.
6. The **TempR-MInt** project, which explores methods by which authorial intent can be integrated into a specialized temporal ordering prediction pipeline. The TempR-MInt work shows potential to assist model design and development within the temporal reasoning space–both for direct improvement of TEO/ETRE and in broader applications of time.

1.1 Research Questions

The driving interest of this research is to define the effects that temporal elements in text have on the attribute of "narrative". Narrative text can exist across multiple domains, genres, and skill levels–much existing literary study of narrative centers on works of fiction written by skilled authors, while NLP finds social media posts by lay writers to be useful in discourse

[1]Repository located at github.com/luke-breitfeller/IlNeS.

analysis (an adjacent area of study). Neither domain is ideal for this dissertation's research. Literature is typically longer than lay writers care to produce; conclusions about literary authors may not generalize to ordinary speakers of a language. Literary analysis is also rarely systematic[2], instead focusing on single texts at a time. By contrast, social media research in NLP surfaces quantifiable trends across broad corpora, but suffers from the non-standard nature of the domain. It is difficult to find social media posts long enough to qualify as a singular narrative, and the interactive element of this medium introduces an external influence which can confound analysis of that narrative[3]. The dissertation requires middle ground: a source of text for which the author's initial intended impact can be isolate, but which still generalizes to the lay population.

This work defines the intended effect of a text as its 'impact'; the impact of a specific text is a function of the author and their goals. For this dissertation to measure impact effectively, the corpus must standardize a short-form narrative format produced by a single author writing without interference. Annotation for the task of temporal event-pair ordering (also known as TEO/ETRE) is difficult, and shaping a new corpus for this type of annotation requires careful attention to the needs of the annotator team. This leads to **Research Questions 1 and 2:**

> **RQ1. How can we effectively annotate new corpora for temporal elements, given the inherent challenges of this domain?**
>
> **RQ2. What can we observe about the link between temporal positioning in text and intended effect from short-form standardized narrative text?**

To answer RQ1, the dissertation contributes a methodological examination of past annotation methods and solutions derived from prior work undertaken during my graduate career. It answers RQ2 with a corpus of standardized short-form narrative and systematic examination of these texts. Conclusions are specific to the concept of "authorial intent" within this dissertation and can be matched to specific types of deviations in temporal order.

As the work builds this systematic approach, it remains concerned with elements of time most useful to NLP work. This dissertation will discuss predictive models for TEO/ETRE, an NLP task which underpins more advanced forms of temporal reasoning. Like most predictive models, TEO/ETRE models exhibit prior class bias in their classification output; a model whose training text has few cases of temporal deviation will replicate that in prediction. It is therefore important to isolate and identify types of text where temporal deviations are more common, so models can adapt. Though there have been past analyses of narrative and time across many domains of text, this particular angle represents a gap in existing research. Therefore, the work asks in **Research Question 3:**

> **RQ3. Are there certain types of narrative which exhibit temporal deviations disproportionate to prior class biases of ordering models?**

To find a standardized format of short-form narrative which can be examined for impact, the work turns to the field of **illness narrative**–autobiographical stories about a personal experience with illness. This restricts the domain the corpus examines to a single topic, which reduces potential confounding factors in analysis, and the choice of illness (or disability) in specific as a topic builds on my past work in clinical NLP. Finally, the field of illness narrative links individual texts to typologies of authorial intent, which makes it ideal for the dissertation's study of time and impact.

[2]Though we do incorporate methods of its approach into this dissertation's work.

[3]A 'thread' of text could be interrupted by the audience, for example, which could alter the temporal order originally intended by the author.

There exist gaps in illness narrative as a field of study; the dissertation responds by building a more rigorous typology that generalizes outside the domain. This supports existing illness narrative work and allows insights from the field to be applied to NLP in a more systematic way. This leads to **Research Question 4:**

> **RQ4. How is the behavior of short-form illness narratives driven by the factor of authorial intent?**

The work produces a corpus (IlNeS) which uses illness narrative as a guiding topic to standardize its short-form narratives. The dissertation also contributes rich analysis of output, to ensure that hypotheses in the field of illness narrative can be grounded in firm data.

Lastly, narrative intention of a text is used within the dissertation to fill gaps in existing TEO/ETRE models. Specialized TEO/ETRE models have previously incorporated specific layers of rhetoric (ex. discourse acts, structural layers). General LLMs, in theory, have the capacity to extract authorial intent from a passage and use it for predictions. Together, these provide decent coverage of rhetoric in TEO/ETRE. What the field lacks is a set of specialized TEO/ETRE models which explicitly capture authorial intent through direct encoding. **Research Questions 5 and 6** are, thus:

> **RQ5. In what ways can authorial intent be incorporated explicitly into existing TEO/ETRE model pipelines?**
>
> **RQ6. Does incorporating authorial intent into TEO/ETRE models improve performance?**

The work contributes multiple experiments which incorporate authorial intent into the MulCo pipeline (a TEO/ETRE model designed for synthesis and co-distillation of information across distinct textual layers). Through this experimentation and analysis, the dissertation examines the ultimate efficacy of this type of information for directly predicting TEO/ETRE.

1.2 Road Map

The dissertation first lays out the foundational elements of the dissertation, including research questions and the road map. It outlines the motivating factors behind its experimental interests and design. As the work exists at an intersection of three existing frameworks (temporal reasoning, narrative analysis, and clinical text), it discusses each element of the intersection as it builds on the last. The dissertation provides deep review of past works in related fields for each of the three broad frameworks, and discusses medicine where the topic relates to text and communication. The work evaluates past annotation efforts within the TEO/ETRE task, highlights challenges of these efforts, and brings insights forward to this dissertation's novel annotation work. It details the design and setup of the illness narrative survey corpus (IlNeS), and analyzes the relationship between the annotated timelines and narrative intention data. It then moves to high-level qualitative examination of testimonials and freeform feedback data to explore trends across the defined illness narrative domain. It details the initial implementation of the TEO/ETRE MulCo model which the dissertation uses as an experimental baseline (discussing performance against existing benchmarks and key qualitative insights). Intention-level information extracted from IlNeS is incorporated into MulCo through the TempR-MInt experimental project and evaluated for TEO/ETRE utility. Finally, the work returns to the driving question of the dissertation–**whether authorial intention correlates with temporal ordering in human text**–answering individual research questions, summarizing the contributions of the dissertation, and exploring potential avenues for future work.

2 Theoretical Framework

This dissertation synthesizes perspectives from three distinct fields of study. These perspectives can be considered to exist in *nested layers* (shown in Figure 1). At the center of the framework is **Temporal Reasoning**, which defines the task of event-pair ordering and discusses prior frameworks for the logic of time. This perspective is layered inside **Narrative Analysis**, from which the work draws techniques to extract the unique intersection between time and narrative. Finally, **Clinical Text** is used as a case study where the temporal-narrative framework can be measured in a single, scoped domain.

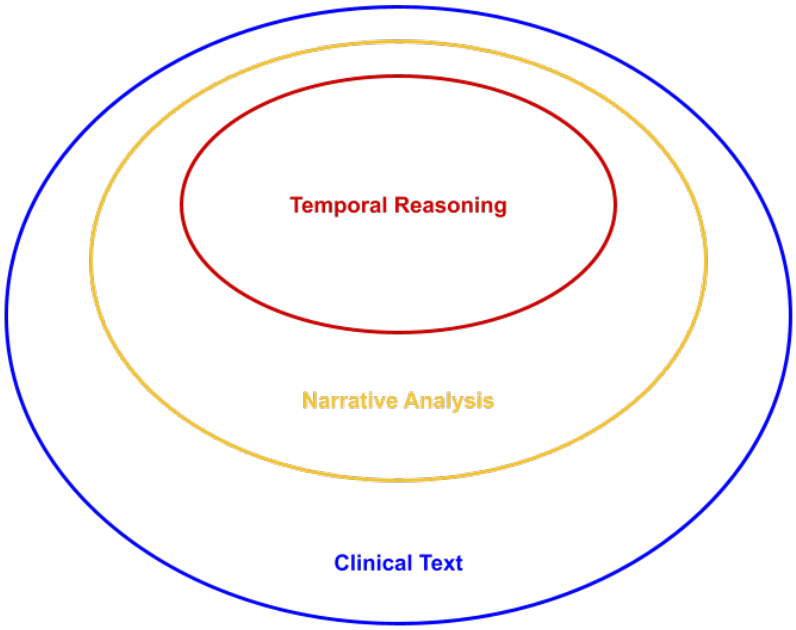

Figure 1: A visualization of the three major inter-layered fields of study in this dissertation.

Each critical perspective presents a unique framework for human text; all are necessary to understand the approach taken within this dissertation. The discussion is grounded with the following text sample (Figure 2):

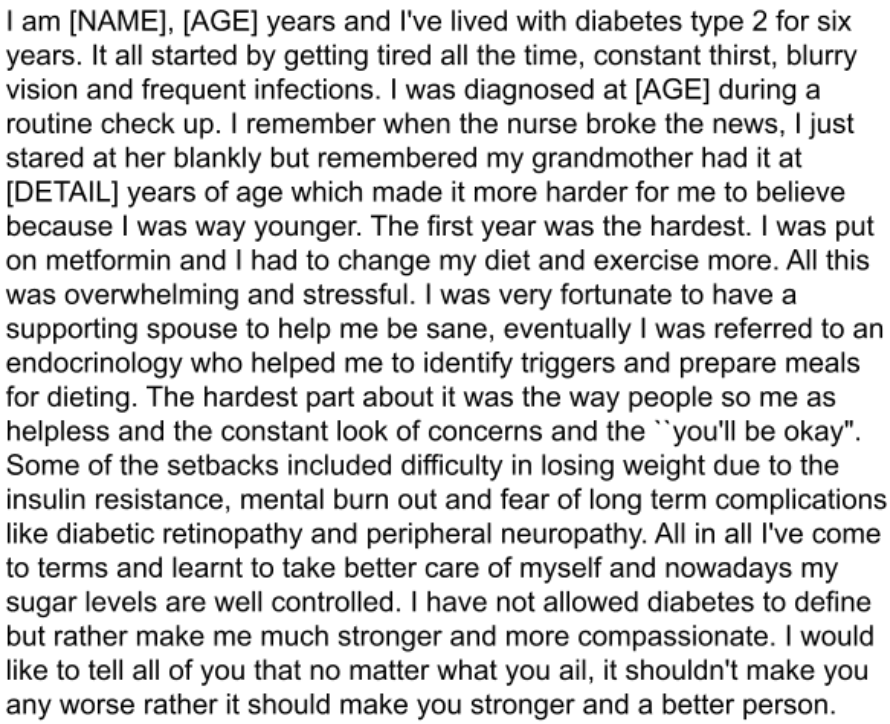

Figure 2: Sample testimony from the dissertation's narrative survey, by Participant 0251.

The text of this sample testimony fits within each of the core disciplines: it references time frequently (ex. "I've lived with diabetes type 2 for six years"), communicates an experience in narrative form, and has a topic which falls within the clinical domain.

The work will return to this example after discussing each critical perspective (**Temporal Reasoning**, **Narrative Analysis**, and **Clinical Text**), to demonstrate how the distinct perspectives contribute to understanding of the text as a whole. As it moves through each, it will explore frameworks where the three primary perspectives form **multi-layered frameworks**.

2.1 Temporal Reasoning

Time is a fundamental property of textual understanding. Therefore, the communication of time is equally fundamental. When one examines the task of temporal reasoning within NLP, there is an obvious, intuitive observation to be made: the clearest way to express time within text is to communicate all events in the same order they occurred in real time.

This is not how most texts are written, which is why temporal reasoning remains a noteworthy task for ML models to solve. As the work defines the specific temporal reasoning this dissertation focuses on, it asks a significant and salient question: *why* do human writers communicate time through disordered narratives?

2.1.1 Task Definition

The research goal of this dissertation is to explore how authorial intent is expressed in narrative on the layer of time. This is done by focusing on **Temporal Event-Pair Ordering** or **Event Temporal Relation Extraction** (abbreviated as **TEO/ETRE**). This task sits within the middle circle of the nested framework ("Temporal Reasoning").

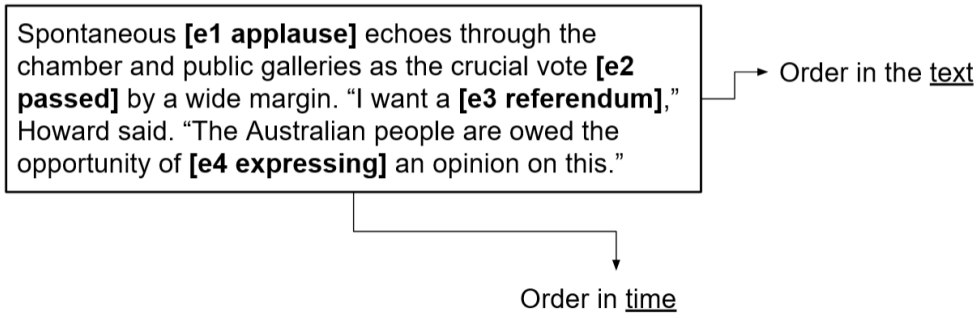

Figure 3: Events within a text.

This dissertation refers to smallest semantic unit used in temporal analysis as an **event**. More specific properties of temporal events will be discussed when relevant throughout the dissertation. Consider Figure 3. Within the displayed text, there are 4 defined events. When the 'order' of these events is discussed, there are two types of order to consider: the order of events within the *text*; and the order of events in *time*.

Within the text, the order of events is $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4$. This may not be their true order in time. The task of TEO/ETRE selects an **event-pair** from the set of events defined. For any potential pair (E_i, E_j) in the set, TEO/ETRE asks in what order the pair occurred in real time, or $Rel(E_i, E_j)$. The types of labels allowed for this operation will be discussed in more detail in Chapter 4.

TEO/ETRE forms the focus of this dissertation because it exists as a foundational building block for more complex temporal reasoning. As an annotation task, it is not without its own challenges (see Chapter 4) and requires that events first be extracted from the original text. But it is foundational for several more complex temporal reasoning tasks (see Figure 4).

- Performing **Timeline Extraction** for a text's 'full' timeline[4] requires a detailed understanding of relative event ordering within a text. The definition of a textual **event** is broad: it potentially covers *all* nouns, verbs, adjectives, and more in a single statement. In practical work, TEO/ETRE restricts its predictions to *relevant* events. This means that even "full" timeline prediction reflects a partial depiction of real-world timelines.
- Many **Question/Answering** tasks may contain a temporal element. These do not explicitly require but may benefit from TEO/ETRE knowledge–for example, the question of "Which actress starred in Luc Besson's *first* science fiction film?" (see Jia et al., 2018).
- **Event Co-Reference**, though not focused on time, is helped by temporal information. Some definitions of co-reference explicitly require temporal similarity (Quine, 1985; Song

[4]A goal of the NIH EpiBio research group, with whom I collaborated during my graduate career and whose work informs this dissertation development (see Section 4.3).

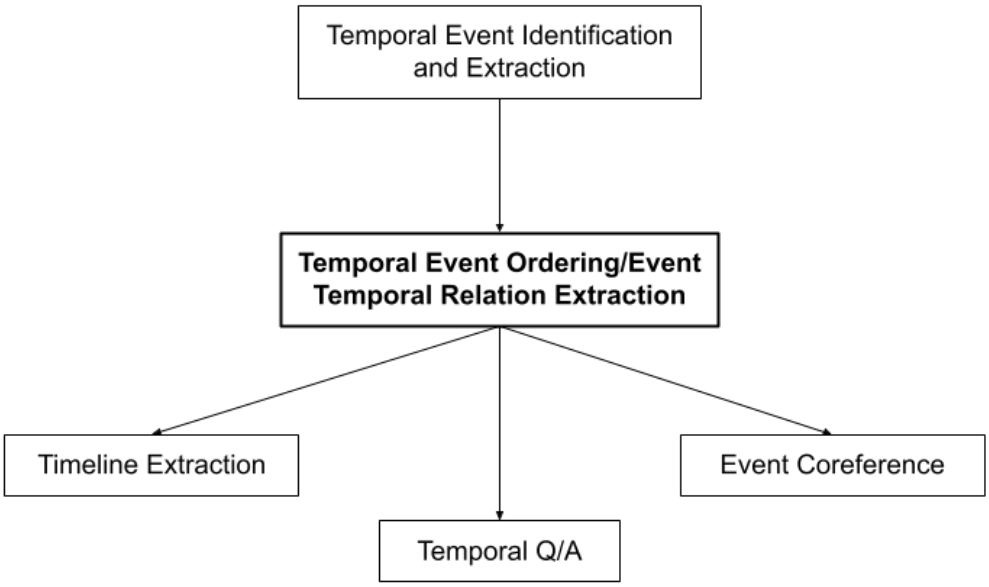

Figure 4: TEO/ETRE within the context of other tasks.

et al., 2015), while others (Davidson, 1969) consider this element 'necessary' but not on their own sufficient.

Note that the inherent limitations of event selection impact the validation process of comprehensive tasks like Timeline Extraction. TEO/ETRE models produce a set of independent predictions of one relation label for each pair within the timeline; even with highly-accurate models, it is expected that some predictions per document will be incorrect. The inherent epistemological difficulty of this work is that a downstream user cannot identify incorrect predictions based only on the model's output. Errors on the event-pair level are difficult to recover from without deep understanding of time and which areas of a timeline might be most prone to model error.

Reliable TEO/ETRE therefore requires a high level of *knowledge*. However, the downstream potential is immense, as many significant NLP tasks benefit from a foundation in time.

2.1.2 Past Frameworks and Models

The necessity for temporal modeling in computational linguistics and NLP have given rise to many frameworks of time hoping to capture the essence of time, as well as ML models which predict for TEO/ETRE in a number of distinct ways. This section discusses these and their impact on the dissertation.

Logical Temporal Models:

In 1983, James Allen produced *Maintaining Knowledge about Temporal Intervals*, which introduced a logical framework of time that sought to comprehensively quantify the ways two

distinct **events** could relate to one another in time (Allen, 1983). This framework defines events using the **time interval** attribute: all events possess a start and end point in time, the span between which is the interval of time for which the event or property is 'true'. Though Allen acknowledges instantaneous events, he finds them irrelevant for the construction of temporal relation typology. It is possible to conceptualize an instantaneous event (a 'time point') as an interval whose start and end are the same, and the logic of the framework remains mostly unaltered.

This definition is useful because it allowed Allen to build a taxonomy of **temporal relations** between events. This taxonomy is fundamental to complex temporal understanding, and key to the work of this dissertation. Allen defines 13 distinct ways in which two events can relate to each other in time, shown in Figure 5 (Allen, 1983). Note that, of the 7 relations appearing, 6 represent *asymmetric* relationships (whose inversions are distinct labels represented within this taxonomy). When examining two events E_1 and E_2, the label that would be assigned to the pair (E_1, E_2) is distinct from the pair (E_2, E_1). The only exception is $E_1 : equals : E_2$, which is fully symmetric. Therefore, the total number of event-pair temporal relations in Allen's framework is 13.

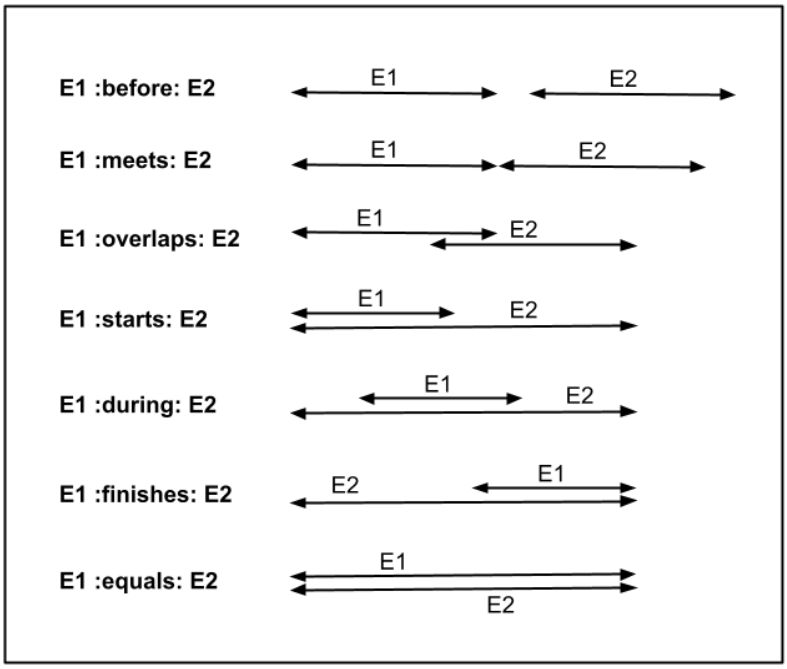

Figure 5: Allen's logical framework of temporal relations, image originally published in Allen (1991). Note that most relation types are *asymmetric*.

This exact taxonomy is often abridged in modeling work[5] but it builds a strong foundation for the task of TEO/ETRE which can be scaffolded by *heuristic knowledge*. When humans make inferences about time, we often reason using 'common knowledge' heuristics. Consider

[5]See UzZaman et al. (2013); Cassidy et al. (2014).

the following:

1. Time moves in one direction, from the past to the future.
2. If two events occur at exactly the same time as one another[6], they must share all **ordered properties** compared to other events.
3. **Ordered properties** possess transitive entailments; if some Event A is before Event B and B is before C, Event A can only be *before* Event C.[7]
4. Cause always precedes (i.e. is *before*) effect.
5. Certain events follow common script formats, where one type of event typically occurs before the other (for example, a cake is traditionally first **baked**, then **iced**, and finally **eaten**).[8]

Early TEO/ETRE (such as Chambers et al., 2014) draw on Allen's logical-heuristic approach to temporal understanding. However, these model types were quickly outpaced by statistical and neural ML architectures (to be discussed further in Chapter 3).

This dissertation nonetheless asserts the value of the human-centered logical temporal perspective. As human readers, we understand the logic inherent to a text's portrayal of time. Therefore, it must be encoded in some layer of the text. Statistical NLP models are skilled at identifying *how* human texts encode information. However, the *why* behind the behaviors of text can often be obscured. Logical approaches to time break down the patterns observed by temporal models, and help researchers identify gaps which can be explained by this underused layer of text.

Statistical Temporal Models:

The 4 standard benchmark datasets used within TEO/ETRE are:

- TimeBankDense: Cassidy et al. (2014)
- MATRES: Ning et al. (2018)
- TDDiscourse: Naik et al. (2019). (This work produced 2 datasets, TDDiscourse-Manual and TDDiscourse-Auto.)

All are built off Pustejovsky et al., 2003b's TimeBank news article corpus. These datasets are discussed in more detail in Chapter 3, with practical applications of the datasets featured in Chapters 4 and 7. Important here is that, when comparing past TEO/ETRE models, a critical distinction is made between **short-distance** and **long-distance** event-pairs.

When TEO/ETRE selects event mentions from a text, these events are drawn from all across the text. Intuitively, a model which seeks to predict for the order between two events should consider the intervening *context* between events. Pairs whose events are closer together in the text may be easier to predict relations for, as the context between events has less noise and contains more direct cues for time. The distinction between short- and long-distance event-pairs is demonstrated in Figure 6.

The traditional cut-off for a short-distance pair has events in the pair appear either in the same sentence or in directly neighboring sentences (commonly expressed as "within 2 sentences").

[6] And only at the same time–that is, there is no point in time for which only one event of the pair is occurring.
[7] See Bramsen et al. (2006).
[8] See Schank et al. (1975).

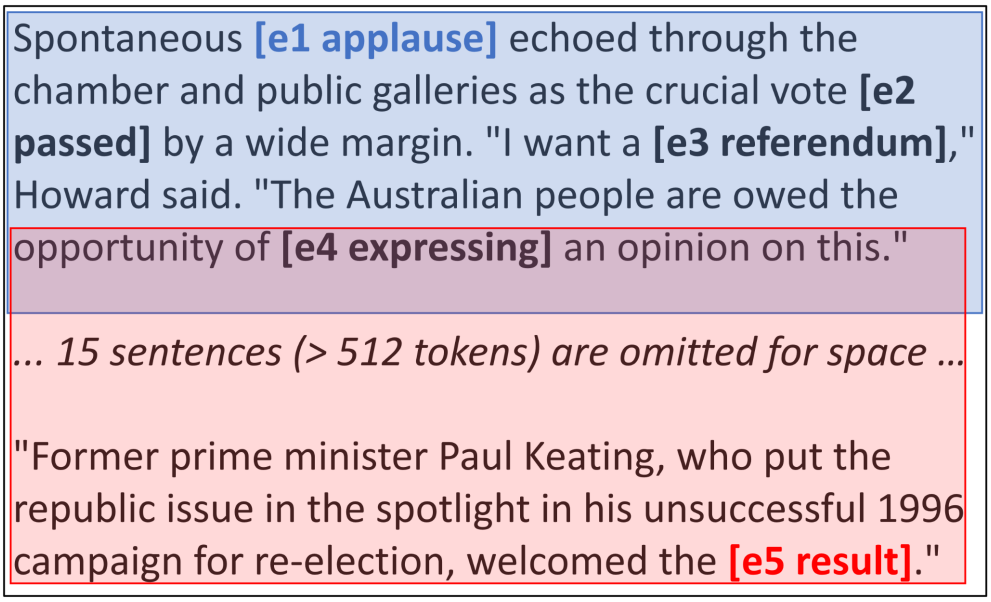

Figure 6: An example of text with both short- (blue) and long-distance (red) context windows. Note the differences in the linguistic cues that are available for each window type.

In the sample, E_1 ("applause") and E_2 ("passed") form a **short-distance** pair. Note that there are significant contextual cues for the pair: "applause echoed [...] as the crucial vote passed"–one can make direct inferences about the temporal relation between these events using context alone. Another short-distance pair (E_1, E_3) requires more implicit reasoning, but could be reliably annotated by humans. Prior to Naik et al. (2019), the standard in TEO/ETRE work was to examine short-distance pairs *only*. This scoped the task for early model work, but enforced inherent limitations on models' predictive power.

Moving to **long-distance** pairs requires a different type of reasoning and modeling. In this example, (E_1, E_5) (where E_5 is "result") is an event-pair with over 15 sentences of distance in the original text. Though it may be possible to deduce the ordering of events given this intervening context, 1) modeling techniques which can encode context for short-distance pairs are less effective for larger context windows, and 2) long-distance pairs require more structural knowledge of the overall text.

As a result, there are roughly three historical approaches to statistical modeling of TEO/ETRE:

1. Direct text encoder models (typically trained on short-distance TEO/ETRE datasets). These include structured perceptrons (Ning et al., 2017) and BiLSTMs (Cheng et al., 2017; Han et al., 2019b). Later models in this and other categories would shift to BERT-based transformer approaches.

2. Structural and graph-based models (which attempt to expand the task to long-distance pairs). This category introduces attention mechanisms to existing models (Beltagy et al., 2020; Kitaev et al., 2020; Zaheer et al., 2020) and GNNs which can better process long-distance dependencies (Liu et al., 2021; Mathur et al., 2021; J. Zhou et al., 2022).

3. Recent SOTA seeks to build on and fine-tune the strengths of both prior methods in

unusual ways. Transformer/graph fusion models, for example, can be improved with careful input data selection (Man et al., 2022) or by re-framing event-pair inputs (Huang et al., 2023).

These approaches, and the models which use them, are discussed in more detail in Chapter 3, along with their comparative performances per each baseline.

Large Language Models:

Though Large Language Models (or LLMs) promise significant improvement over smaller models on a variety of tasks, benchmark studies for temporal reasoning demonstrate gaps in these tools' performance. These studies are discussed in more detail in Chapter 3–here, this work notes that three key benchmark studies (Chu et al., 2023; Y. Wang et al., 2023; Wei et al., 2025) show successes with LLMs on many forms of temporal reasoning, but see under-performance compared to human annotators and other ML models on tasks which require *symbolic and implicit reasoning*. For Y. Wang et al. (2023)'s TRAM benchmark Relation task (which predict the same event-pair ordering relationship as TEO/ETRE), all LLMs in the benchmark study performed worse than simpler BERT/RoBERTa models.

The comparison of these LLM results to smaller, more focused models demonstrates that *model size alone* does not predict performance for this type of reasoning. The authors cite "several factors" that might explain their result: "RoBERTa-large[9] may utilize optimization strategies particularly beneficial for these tasks. Additionally, inherent features or efficiencies in its architecture might enhance its ability to understand and process temporal cues" (Y. Wang et al., 2023).

This observation builds a key motivation for the dissertation: **What are the architectural efficiencies present in 'smaller' language models made evident by these benchmarks, and what can they tell us about modeling time?** For tasks where human annotators outpace statistical models, the dissertation hypothesizes the human element is what will shape successful machine models.

2.1.3 Sample Application

With this understanding of past works in the field of temporal study, the perspective of time can be applied to the sample text (Figure 7, below).

This example highlights linguistic cues of the type used to communicate time within text. Temporal cues can be separated into distinct topics. These topics carry semantic meaning, as the events in the illness journey are broken into meaningfully-distinct segments of the overall experience.

1. **Pre-diagnosis**: the symptoms which led to diagnosis.
2. **Diagnosis**: the specific appointment.
3. **The first year**: initial treatment and lifestyle changes.
4. **'Eventual' referral**: the referral to an endocrinologist and new treatment plans.
5. **Current day**: a stable equilibrium.

[9]The most successful of the BERT/RoBERTa models tested for the task, but significantly smaller than the LLMs included by the study.

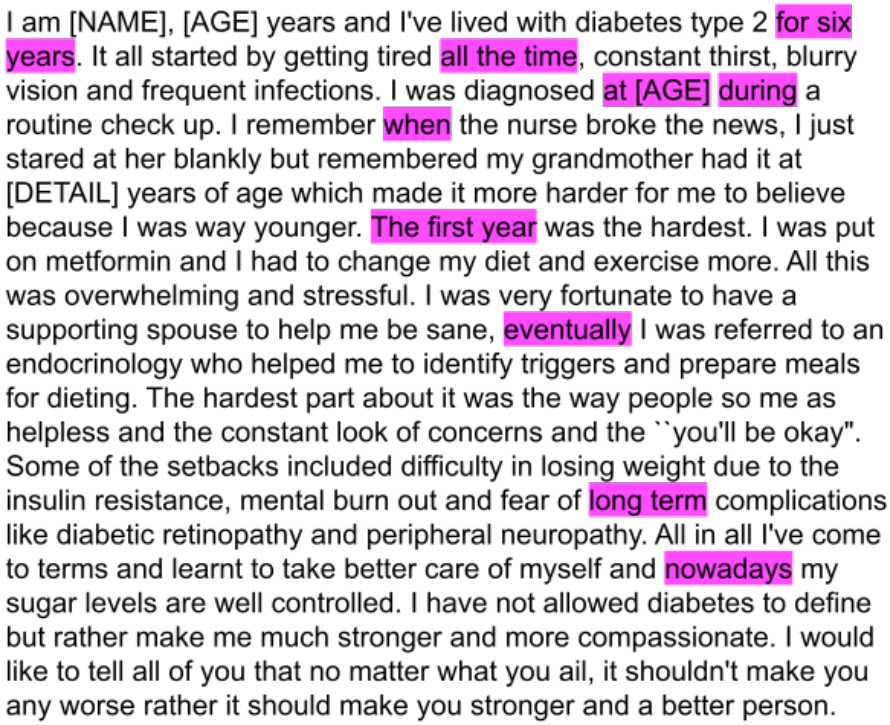

Figure 7: Sample text, annotated for the time layer (in pink).

The passage follows a nearly-chronological ordering. There is one temporal linguistic cue ("for six years"), which covers the entirety of the survey participant's described experience, and another which references hypothetical future events ("long term complications"). These exceptions to chronological order represent potential sources of temporal deviation that would need to be captured by a TEO/ETRE model.

ML models can be trained to recognize and use temporal cues like these to build relation predictions. Encoder models can take in the cues directly using context windows, while GNN models might affix relation cues into a semantic graph. Structural elements that hint to the general chronological ordering can be learned by long-form ML models, and LLMs are capable of some temporal reasoning with long input text.

This sample text shows how the axis of time can **build structure** within communication. More than a simple description of events, time becomes associated with semantic meaning within text. As the work moves to narrative analysis, it explores how that semantic meaning is intertwined with deeper emotional aspects of narrative.

2.2 Narrative Analysis

This section moves up in the framework from the perspective of "Time" to that of "Time-and-Narrative". Narrative exists as a subject of study within NLP, but this dissertation also incorporates perspective from the humanities. It argues that the narrative attribute of a text is

not a fully separate layer from time, and that information about one can inform the other. This section discusses existing theory supporting that argument, examines the "Time-and-Narrative" perspective, and shows how this fusion of approaches can answer dissertation questions.

> "The narrative **is a temporal sequence**. A doubly temporal sequence, one must hasten to specify: There is the time of the thing told and the time of the telling (the time of the significate and the time of the signifier)."
>
> -*Film Language: A Semiotics of the Cinema* (Metz, 1991)[10]

As an element of text, narrative is posited by theorists to exist as a distinct attribute texts may or may not possess. Texts which describe facts, instruction manuals, and even creative work such as lyrical poetry could all fall outside the bounds of what is considered *narrative text*. Gérard Genette defines it as the "the signifier, statement, discourse or narrative text itself" surrounding a story–that is, the means by which the story is told (Genette, 1980). He presents a theoretical triangle relating its primary components as **story** (content of message), **narrating** (delivery of message), and **discourse** (the method by which the message is packaged to an audience), with elements of text itself building relationships between the three. This is similar to the *structural* perspective of language, where elements of text build meaning only through interaction with one another.

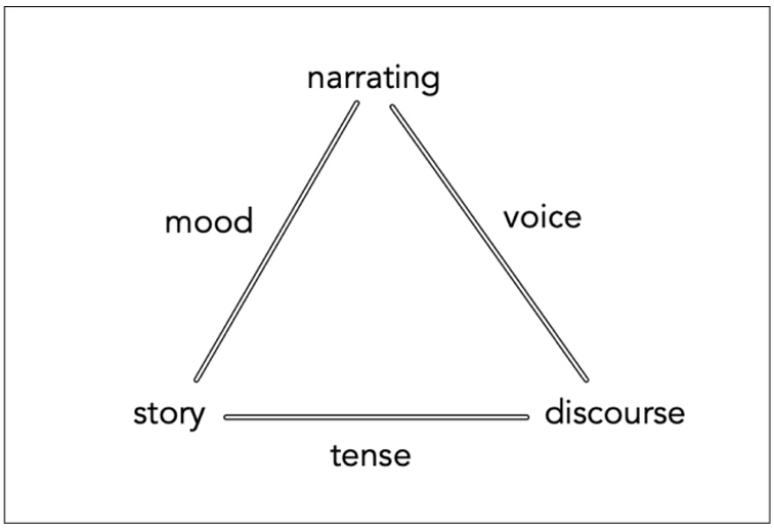

Figure 8: Genette's narrative triangle, image produced by Piper et al. (2021).

But the narrative triangle does not necessarily differentiate texts which are truly "narrative" from those which merely *describe* events. Piper et al. (2022) filled this gap by collecting human annotation of various texts, having annotators rank them for their perceived narrativity. Linguistic analysis was used to surface trends across narrative texts. The work found correlations between narrative text and the attributes of "experientiality", "world-making", and "sequentiality"–in other words, a narrative must by definition capture a sense of agent, world, and time (Piper et al., 2022). This last element is key to the dissertation: the framing of narrative as a subset of all human text defined by its *inherent interactions with time*.

[10]Original text published in 1971.

The temporal meaning communicated by a narrative, in short, is not incidental. Narrating, as a means of information delivery, introduces delay to text. Few stories about events are told at the same real-time moment as their events occur. Further, readers will take in a text at somewhat-constant reading speed, but real events may have occurred for distinct durations in real-time. The temporal meaning communicated by text is fundamentally distinct from the experience of reading it, but human understanding of narrative expects sequentiality to be encoded in the discourse layer. To quote Metz (1991) again, "One of the functions of narrative is to invent one time scheme in terms of another time scheme." Likewise Labov et al. (1997) describe narrative as: "a particular way of recounting past events, by matching the order of narrative clauses with the original order in which those events occurred. Thus narrative, as defined by 'temporal juncture', **is only one of many ways** of dealing with the past."

Narrative analytical technique allows the work to surface the **impact** of a text, and the effect that temporal ordering in specific has on that impact. The theory and frameworks introduced in this section argue that this effect is not a simple consequence of how the text is constructed, but an intentional effort on the part of the text's author (whether deliberate or subconscious). Note that here, *intentional* is used as a contrast to *accidental*. This work argues that the effect of temporal ordering on narrative impact is understood across one's linguistic culture, and that even lay writers unconsciously build this understanding with other communicative knowledge. Therefore, even when a lay author does not understand how the temporal deviations they introduce to a text change its narrative, that behavior is still meaningful to the overarching research question; it may be an unconscious reflection of broader cultural trends, but it is not random.

2.2.1 Time Impacts Narrative

Narrative is an organizing element of text, producing an understandable through-line which resonates with readers. Time exists within text for a reason, and the frequency of "temporal distortions" present within text shapes the final impact of the narrative as much as rhetorical elements like register, genre, and individual style.

> "It is generally agreed that a narrative is the presentation in discourse of a sequence of past events [...] it is also agreed that this sequence does not begin or end by chance, but has a beginning, a middle and an end recognized **by some principle**, so that the listener knows when a narrative has begun and when it has ended."
>
> -*The Language of Life and Death: The Transformation of Experience in Oral Narrative* (Labov, 2013)

This discussion of sequence in narrative presents events as having roughly chronological sequence (moving from beginning, middle, and end). But narratives are not exclusively chronological; temporal deviations in text do not simply change the duration of events in time (as discussed by Metz), but also their order.

In literary analysis, singular texts are examined and the effects of their technique articulated and explored. Here, this dissertation will do the same. Literature is a genre of long-form text which conveys a sequence of (often, but not exclusively) fictional events. They do not match to a "ground-truth" sequence of events outside the text, nor are they designed to educate readers about the sequence of fictional events. Therefore, the primary purpose of a work of literature can be said to evoke **mood** and **experience**. By examining the order of events within a work of literature, conclusions can be drawn on how this ordering alters mood.

Consider the case of Bram Stoker's 1897 novel *Dracula*. *Dracula* is a gothic horror[11] novel written in epistolary format, covering journal entries discussing the events of the book and letters exchanged between characters over the course of the (fictional) narrative. These epistles overlap in "real" time, but in Stoker's arrangement, segments of the book center around specific narrators' point-of-view, even when this requires skipping around the timeline.

The largest 'jump' in the original text moves from the point-of-view of Jonathan Harker, trapped in Dracula's estate and fearing his oncoming death, to a months-earlier exchange of pleasant letters between characters Mina and Lucy. This choice has been directly cited as having strong impact on the mood of the text:

> "There is no greater jolt in Dracula than this **abrupt transition** from horror to domestic happiness in Lucy's letters from Mina. It is as if the novel has changed mood without sacrificing immediacy."
>
> -*The Narrative Method of Dracula* (Seed, 1985)

Because *Dracula* is an old text within the public domain, the ordering and mood of the original can be contrasted against derivative, transformative works. For example, *Dracula Daily*, assembled by Kirkland (2021). *Dracula Daily* is an email subscription blog hosted on Substack blog, which mailed subscribers excerpts of the text on the same month and date as they had been "written" within the fictional story. It is, in short, a chronological retelling of the tale. Readers of the blog reported distinct feelings when experiencing the book in this format compared to the (non-chronological) original text. Quote one:

> "I think it's amazing to have a bunch of readers who are reading this book—not as Bram Stoker wrote it—but **in a way that conforms to the steady march of events within it**. [...] You can't have your dread or anticipation undercut by future events. Like all the characters you're going to meet, you just have to wait for Dracula to act upon you."
>
> -Tumblr post (atundratoadstool, 2022)

As a transformative work, *Dracula Daily* serves as an ideal case study for the way **temporal ordering** impacts **audience response**. The only element of the text which has changed from the original is the order in which events were presented to readers. Though both versions evoke fear, one "jolts" between moods and maintains tension through "immediacy", while the other slowly builds a helpless sense of dread. These nuances are not insignificant; they showcase the power of temporal ordering to evoke emotion within narrative.

This establishes a connection between narrative time and audience *reception*. But humans do not just interpret natural text–they create it. Literature (particularly works considered to be part of a society's 'canon') represents a high level of *narrative skill* on the part of the author. The "abrupt jolt" Seed identifies in *Dracula* was likely deliberate on Bram Stoker's part. The dissertation contends that these motivations are not reliant on high narrative skill; these techniques are used by even lay human authors, whether or not they remain consciously aware of that fact.

Building on the question raised in Section 2.1, this work hypothesizes the following explanation for disordered narrative in human text: **authors present temporal information out of order when it benefits the narrative to do so**. By applying techniques from narrative theory (close reads of text, and the re-situating of text in its original context) to texts with

[11] "Gothic" and "horror" both representing sub-genres associated with the mood of *fear*.

temporal deviation, it is possible to extract this interaction between **narrative** and **time**. From there, the work builds a framework to show how both internal and external motivating factors can influence even the lay creation of text.

2.2.2 Close and Deep Reading:

In the digital age, humans are able to interact with texts in ways that had not been previously possible; reading and writing text has become more accessible, new forms of media like web pages utilize text in unique ways, and tools like ML models are able to process text at scale. Paradigm shifts like these open new avenues for textual analysis–and yet, many are concerned that the digital era may cause essential older skills for text analysis to fall by the wayside. UCLA Professor Maryanne Wolf notes, in a piece advocating for deep reading practices of the type used in the humanities:

> "An early immersion in reading that is largely online tends to reward certain cognitive skills, such as multitasking, and habituate the learner to immediate information gathering and quick attention shifts, rather than to deep reflection and original thought [...] the digital culture's reinforcement of rapid attentional shifts and multiple sources of distraction can short-circuit the development of the slower, more cognitively demanding comprehension processes that go into the formation of deep reading and deep thinking."
>
> -*The Importance of Deep Reading* (Wolf et al., 2009)

Though Wolf statement focuses on the preservation of deep reading techniques, the skills encouraged by digital reading are not lesser to the skills of deep reading–they are simply different methods of understanding and synthesizing information about text. But that distinction forms an important motivation for this dissertation work; though all methods of approaching a text are *valuable*, they are not *interchangeable*, and many NLP tasks respond best to distinct forms of analysis.

Machine learning in NLP often takes a **bottom-up approach** to textual processing: statistical models take in large input corpora containing specific examples of linguistic phenomena to learn general patterns across the medium. Technology in the field continues to improve, allowing for larger and larger training sets, which means a **top-down approach** of the type generally practiced in literary and narrative studies becomes increasingly infeasible. Yet, there is value in the top-down, deep reading model. The most common NLP framework for predictive tasks presents a pipeline whose input is primarily (if not solely) text. It is a known fact that information external to a text (word definitions, cultural context, and so forth) can improve ML understanding of the text itself, but passing that information to a model is often challenging. When models incorporate external context[12], it is often as an addition to existing model architecture (see Figure 9)–through metadata, knowledge graphs, and similar encoding methods. ML-driven NLP also pursues context with other **bottom-up** techniques: from the introduction of bag-of-words to current LLMs, NLP makes the assumption that by drawing connections between texts in large dataset, a model's architecture can infer context that would otherwise be invisible.

However, these methods do come at cost–knowledge graphs, for example, are notorious for requiring expert annotation at large scale to provide effective external context to models. Though

[12]Through the remainder of Section 2.2, 'context' refers to the *elements external to a text*. This is distinct from the 'context' used by context-encoder models, which define *excerpts of a text* which are not the direct input of a predictive model but which still inform output.

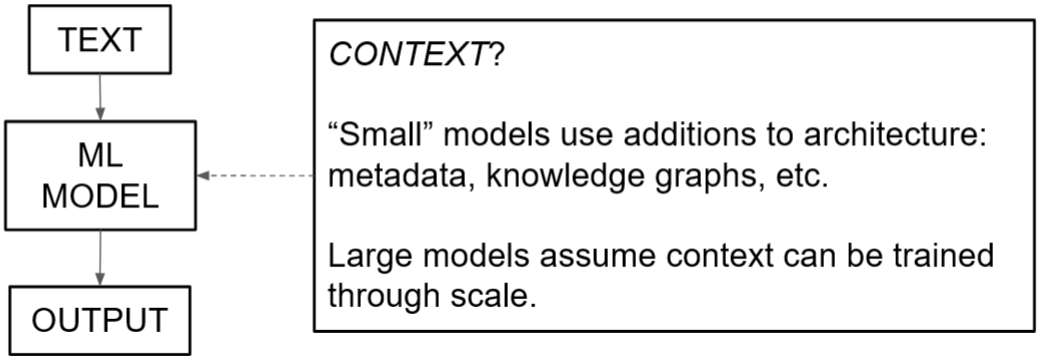

Figure 9: The basic NLP pipeline.

LLMs present themselves as low-annotation methods for encoding data, these models in fact build off of extensive annotation and fine-tuning efforts (Heikkilä, 2023; Dinika, 2024). The introduction of narrative analysis to the perspective of NLP poses a significant question: **is the bottom-up framework truly the best approach across all tasks for which we apply natural language processing**? If not, what types of task could benefit from additional external context?

Deep and close readings of individual texts are associated with the humanities. Researchers in these fields examine texts on the level of sentence, phrase, and even specific words to tease out meaning. They also and consider texts in the context of greater historical and societal trends. In the pedagogic work *Teaching Literature*, Showalter defines “close” reading as “[...] a deliberate attempt to detach ourselves from the magical power of story-telling and pay attention to language, imagery, allusion, intertextuality, syntax, and form” (Showalter, 2002). Wolf defines the related concept of deep reading as “the array of sophisticated processes that propel comprehension and that include inferential and deductive reasoning, analogical skills, critical analysis, reflection, and insight”, and emphasizes original thought as a key element of a responsive, deep reading (Wolf et al., 2009).

Deep and close reads of text are, as noted, typically not feasible in NLP work due to the scale of data that ML models require. And yet, areas of language with which NLP struggles may benefit from deep, qualitative evaluation. Showalter’s citation of deep reading as a detachment from the “magical” experience of a story describes a process similar to how interpretability methods are applied to opaque ML models. In both, the purpose of examination is to illuminate the mechanisms that make a system (be that a story or a machine-learning model) operate. This dissertation collects new text for analysis and seeks close readings of those texts to help explain the mechanism of narrative impact. Though the work does aim to present quantitative analyses of the corpora, it also explores the qualitative insights that careful reading of these texts (both in isolation and in the wider context of illness narrative) can bring to the ML space.

2.2.3 Text and Context

Machine learning frameworks often treat *texts* as fully self-contained objects. *Context* could be interpreted as types of knowledge regarding an object (this definition of context is distinct

from the 'context between event-pairs'[13] which is encoded into TEO/ETRE model), but the object itself is an isolated system.

Figure 10: In text-as-object framework, the text is interpreted as a real-world object.

In the representational metaphor illustrated by Figure 10, when NLP models take a text as input, they measure the shape of the object (here visualized as a white vase with notable features like color, height, and shape of handle) directly. Contextual elements like metadata or knowledge graphs provide the model with knowledge of what shapes correspond to what objects. From there, the model can make confident predictions about the nature of the object through a primarily textual analysis.

This perspective is not without philosophical grounding: *structural linguistics* as a field holds that language primarily functions as a self-contained system, which can only be analyzed and understood through reference to other elements of language. Modes of literary analysis building upon the structural (like the New Critics) advocate for analyses of text which ignore authorial intention, cultural context, and reader response in favor of close reading the text alone. And yet, structural philosophies of language hint at an inherent limitation of text-only focus: structuralism posits language as something that is divorced from physical, ground-truth reality, and separate from the underlying thought.[14] This perspective is not feasible within NLP for the simple reason that **many models are asked to make predictions for both language and thought**.

Returning to the metaphor from Figure 10, a structuralist philosophy of language cannot consider text to be a real object. At best, it is a reflection or shadow of the object (see Figure 11). There may be identifying features in the reflection which can help to deduce the real, original object (ex. general shape, the handle), but others (color, height) have been lost in the translation. The shadow itself will always be flattened compared to the original.

This metaphor draws intentional similarity to the thought experiment of Plato's allegory of the cave. Plato's cave follows a theoretical population living inside a cave, who can only experience the world through shadows cast on a wall. As these people cannot contextualize a reality beyond their perception, they would come to believe that these shadows were all that existed in the world. An ML model, likewise, only has access to what it is provided as input. For text-based models, this is the text itself and certain kinds of context. It is possible to conceive

[13]Within this chapter, *context* will be used to mean elements outside a text which nonetheless may inform the text. Context between event-pairs, in this framework, is still part of the original *text* object.

[14]The relationship between language and thought has been explored from the earliest days of psychology as a field. Lev Vygotsky presents such a framework in *Language and Thought* (Vygotsky, 1986) (first published in 1934). Vygotksy contends that communicative language is not a perfect proxy for thought, and yet a link between the two is undeniable.

Figure 11: In text-as-shadow framework, text is a projection of ground-truth reality, and lacks dimension compared to the original.

of ML models (even Large Language Models) as cave-dwellers in some sense. They do not have access to ground-truth reality–despite often being asked to make predictions about that reality itself.

Consider two different types of NLP tasks. **Sentiment analysis** is a task in which a model receives a text and must judge the sentiment present within the text. Though there are contextual elements to our understanding of 'positive' and 'negative' sentiment, the model is being asked a question about the *text itself*. In our metaphorical framework, the model is using the shadow to judge the shadow (Figure 12).

Figure 12: In Sentiment Analysis, the model is given a text and asked to make predictions about the text.

But in temporal tasks using text which describes real-world events (ex. **Temporal QA**), the model is being asked to judge on *ground-truth reality*. It is using the shadow to predict for the object. This dissertation argues this metaphor captures an essential aspect of textual descriptions of reality, demonstrated in Figure 11: they lack dimension compared to the real. The process of communicating real events in text is akin to lossy compression–authors may forget details that would aid a predictive model, or they may withhold these details on purpose. The process of converting reality to a textual representation, therefore, is not completely reversible. The text is a shadow, cast upon a wall.

Theoretical Framework:

How then, does NLP move from text to reality, if the 'casting' of a shadow cannot be reversed to produce the original? Here, the work moves back to the metaphor and uses it to produce a framework that can be applied to text. The allegory of the cave, as a thought experiment, is founded on a basic philosophical axiom: human conception of reality is driven by what we are capable of perceiving directly. The concern of the dissertation is *how* and *by what factors* that perception can be shaped. To that end, the metaphor will be somewhat loose–what is

important is that this example outlines the factors which come together to produce the 'reality' reflected in a text.

One can think of the concept of shadows on a cave being built from five parts, shown in Figure 13: 1) the original object, 2) the surface it is projected upon, 3) the light source, 4) the observer, and 5) the shadow that results. These will be discussed in order of their simplicity in the metaphor.

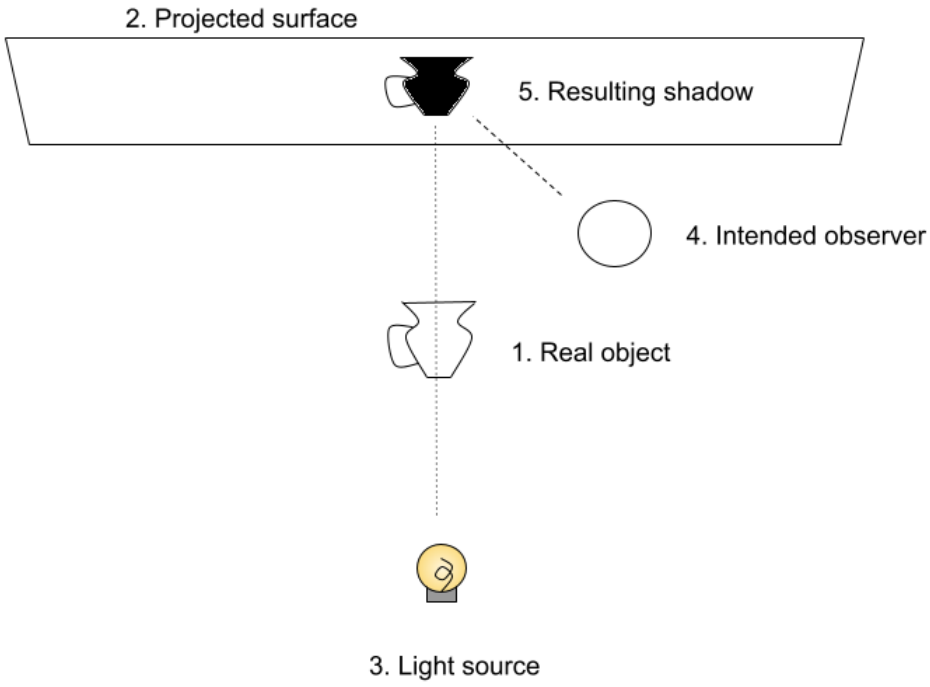

Figure 13: Illustration of the full shadowed cave metaphor.

The **original object** (1) symbolizes ground truth, the reality that is being described within a text. In many text domains (ex. news articles, clinical documents, nonfiction), the events which happened in reality are fully outside the control of the individual writing the text.

The **resulting shadow** (5) represents the text itself. It is what is seen by the observer, who is unaware of what exactly the original object looks like. Natural language processing seeks to step backwards from the text to the object itself; the dissertation argues that this process often excludes important additional factors.

The **surface of projection** (2) can be said to represent external factors that limit the ability to project the text. In the metaphor of shadows on walls, a shadow cast on a perfectly flat wall can be made to look much like the object it represents. But if the surface itself has bumps and perturbations, it may be impossible to produce any resemblance to the original object.

External factors like ordinary memory deterioration, decrease in mental function due to illness or stress, and loss of records can influence the construction of a human-generated text. These are not stylistic choices, as they are outside the control of the writer, but their impacts may

appear similar to those of register, genre, and style.

The **observer** (4) represents the audience of a work. In the metaphor of a shadowed cave, the audience has no control over the object that is being projected, the cave wall, or the light that is casting the shadow. And yet, their position in the cave can influence how the shadow appears; an audience looking at the shadow head-on may see something different from an audience looking at it from an angle. They are not changing the shadow itself, but the work discusses below how their presence can influence the creation of the shadow in indirect ways.

When NLP researchers study the process of communication, one aim is to identify what knowledge, emotion, or other concept is being passed along from the original author of a text to their readers. But just as those in the cave cannot see the original object, NLP often lacks access to that audience and their perspective. Researchers are forced to make inferences about who might be reading a text and why, what the intended audience of a work is and if the actual audience differs. Annotation labels take the perspective of an average, reasonable audience member, and those are used to approximate the end result of the lines of communication.

The central object in this metaphor, and the one this dissertation considers most significant, is the **light that casts the shadow** (3). Within the metaphor, imagine that the light source is a candle or flashlight that can be carried and manipulated by someone in the cave. That person cannot alter or move the object they are casting a shadow of, they cannot change the cave wall, and they cannot guarantee that their audience will sit where they want them to. All they directly control is the light.

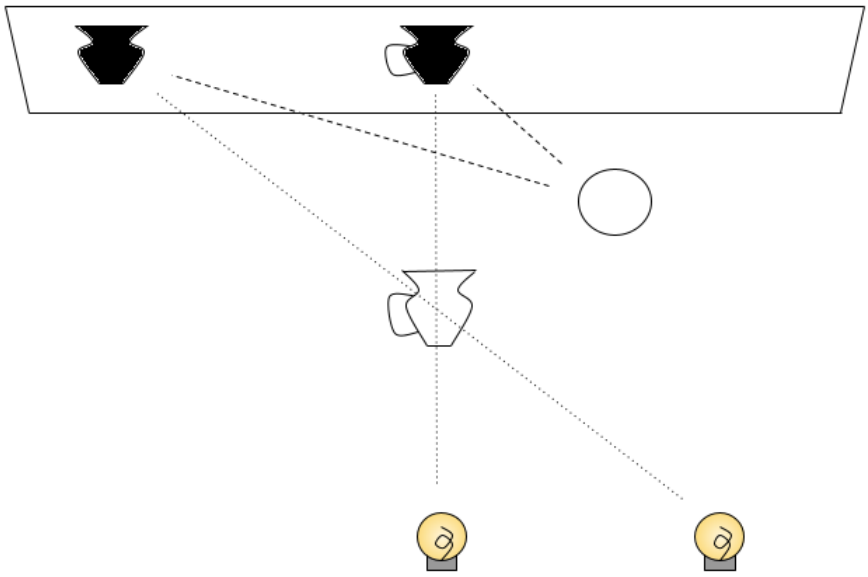

Figure 14: Repositioning the light source may significantly change the visible shadow on the wall.

Under those constraints, the shadows this person can cast may still vary. Some features of the original object may be prominent or obscured depending on the angle of the light (consider the

handle on the side of the vase in the example). Different sections of the cave wall may affect the shadow. Finally, the audience member may see the shadow differently, depending on where they are sitting to view it. In this example the person holding the light may choose to move their light to ensure the shadow is cast a certain way in response to any of these features (see Figure 14). They may choose to highlight features of their original object, avoid a section of the cave wall they know will distort the shadow, or project the shadow in a way that presents a better image based on the expected audience and their position in the cave. The person holding the light cannot *change* these outside factors, but they can *respond* to each of them.

How this individual responds, further, reflects an influence unique to this piece of the metaphor and critical to this dissertation: the individual's intent in casting the shadow. Is the person holding the light trying to make it clear and obvious what the original object is? Are they trying to frighten their companion with a large and dramatic shadow? Is there simply an interesting shape that can be made with the shadow cast at the correct angle? Each intention will lead to a different shadow, one that is nonetheless influenced by ground-truth reality, external distortions, and the perspective of the observer.

In this text, this attribute is defined as **authorial intent** or **impact**: what it is the author intends to communicate through a given text.

2.2.4 Sample Application

Returning to the sample text, the narrative lens is added to the temporal analysis:

I am [NAME], [AGE] years and I've lived with diabetes type 2 for six years. It all started by getting tired all the time, constant thirst, blurry vision and frequent infections. I was diagnosed at [AGE] during a routine check up. I remember when the nurse broke the news, I just stared at her blankly but remembered my grandmother had it at [DETAIL] years of age which made it more harder for me to believe because I was way younger. The first year was the hardest. I was put on metformin and I had to change my diet and exercise more. All this was overwhelming and stressful. I was very fortunate to have a supporting spouse to help me be sane, eventually I was referred to an endocrinology who helped me to identify triggers and prepare meals for dieting. The hardest part about it was the way people so me as helpless and the constant look of concerns and the ``you'll be okay". Some of the setbacks included difficulty in losing weight due to the insulin resistance, mental burn out and fear of long term complications like diabetic retinopathy and peripheral neuropathy. All in all I've come to terms and learnt to take better care of myself and nowadays my sugar levels are well controlled. I have not allowed diabetes to define but rather make me much stronger and more compassionate. I would like to tell all of you that no matter what you ail, it shouldn't make you any worse rather it should make you stronger and a better person.

Figure 15: Sample text, annotated for the time layer (in pink) and narrative layer (yellow).

This text (shown in Figure 15) belongs to the sub-genre of illness narrative (to be discussed more in Section 2.3.3). This sub-genre does not have a standard register. Yet, the author has selected one that is relatively formal. There is little slang (slang might suggest the author is aiming to create a casual narrative experience), but also only uses medical jargon with purpose. This suggests the author is seeking to appear neutral and informed to their reader, without presenting more expertise than they genuinely have.

It is possible to identify framing statements within this text, which suggest a structure to the underlying narrative. The author begins by introducing themself and ends with a direct address to the 'reader'. Following the beats of the text, there is rising action ("It all started when" → "The hardest part about it" → "setbacks") followed by a resolution and sense of completion in the narrative ("I've come to terms"). Notably, though this author has mostly been speaking of their experience in factual terms, they end their text with a lesson for the reader. "It shouldn't make you any worse rather it should make you stronger and a better person." This statement, though outside the actual chronology of the author's experience, clearly serves a narrative purpose in tying their experiences together. The events do not simply *end* but *conclude*, in a narratively satisfying way.

When we examined the sample text only through the lens of time, we identified 5 temporal periods in the text: 1) pre-diagnosis; 2) diagnosis; 3) the first year; 4) eventual referral; 5) current day. But by examining how these events are discussed narratively, it is possible to build a more thorough understanding of how the participant perceives and reacted to the original events:

1. **Pre-diagnosis**: Pre-diagnosis covers a short segment within the original text. Though symptoms are numerous, discussion of them is brief.
2. **Diagnosis**: Diagnosis reflects a single moment temporally, but covers multiple sentences in the text. The participant referenced specific details of their memories of the moment and related it to their personal history and emotional state.
3. **The first year**: The first year of treatment is marked specifically as a difficult time, and the author provides details about their emotional state and resources which helped them through their struggles.
4. **'Eventual' referral**: The use of the word 'eventually' evokes a slow progression. It is the last mention of time within the passage until much later, giving the events after the referral a somewhat ephemeral quality.
5. **Current day**: The present moment is only referenced at the very end of the text, as part of the final 'moral lesson' of the piece.

Like all of the work's survey participants[15], this participant was not required to have a skill level in written English above comfortable fluency. Though some expressed higher skill levels, the average participant appeared to have only lay writing skill. But laypeople are constantly exposed to the standard of language used within their culture, and they replicate (consciously or otherwise) these standards in their own texts. Studying the language of average writers with an eye towards time as a tool of narrative, therefore, reveals natural tendencies within language groups to tailor temporal orderings in text for specific narrative purpose. Those tendencies are examined in more detail in Chapter 5. Next, the work moves to the clinical domain of text.

[15]Recruitment is discussed in detail in Chapter 5.

2.3 Clinical Text

Time has always mattered in clinical understandings of illness. As one example, *pain* is an expected part of healthy human experience. But *persistent pain* which recurs for 3 months may be diagnosed as a chronic medical condition[16]; distinguishing between these two conditions requires an understanding of time. Temporal analysis of illness aid in many clinical tasks–for example: diagnosis, prognosis, and epidemiology. Much of clinical temporal NLP deals with **clinical records**, texts made by clinicians about interactions with a patient as part of standard healthcare documentation. Most clinical records represent day-of-appointment observations (though clinical records in other formats do exist).

One reason why clinical records are the preferred document type for clinical NLP is scale. New clinical records are produced daily, and represent patients across every possible demographic. However, there are drawbacks to working strictly with these texts. Their focus is on a narrow window of a patient's experience (namely, the days in which they meet with clinicians for appointments). But patients live with illness consistently across their day-to-day life. To get a thorough understanding of a patient's condition, it is often necessary to extract patient-level self-reports from clinical and other texts–which often mirror narrative texts in construction and intent.

For example, temporal cues can be linked with illness phrases from the sample testimony used within this chapter to produce the timeline in Figure 16.

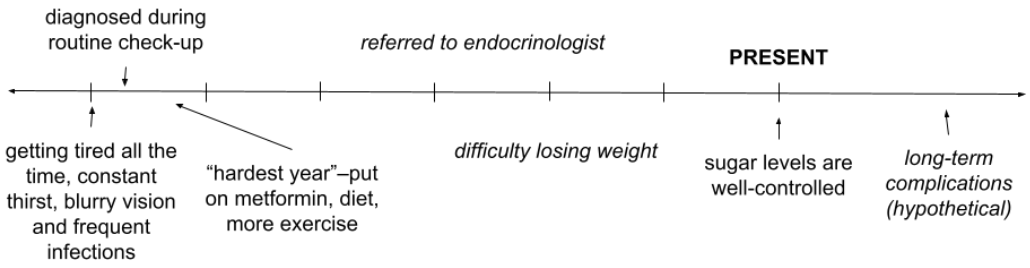

Figure 16: The symptoms reported in sample testimony, arranged on a timeline.

The clearest information discusses the onset of this condition (six years ago). The start of symptoms, diagnosis, and first year of treatment are anchored with direct temporal values. There are two events which cannot fully be placed on the timeline–"referred to endocrinologist" and "difficulty losing weight". Readers do not know in what year these event occurred; all information is *relative* to other events. The endocrinologist referral, for example, occurred at some point after the first year of treatment, but before the present moment where sugar levels are "controlled".

The ambiguous quality of textual timelines explains why TEO/ETRE is often a more useful tool than strict date extraction for clinical temporal analysis. In many cases, particularly in less-formalized genres of text, useful clinical events are not associated with a date. Chapter 4 discusses further how TEO/ETRE can be leveraged to produce clinically-useful timelines from freeform text. Here, the dissertation builds a case for how patient self-reports (either embedded within clinical texts or sourced from other records) follow the temporal and narrative framework laid out in prior sections of the work.

[16]See the International Classification of Diseases (WHO, 2022).

Note about verbiage: in NIH research, texts within the medical domain are referred to as 'clinical text' and expert authors in the domain (whether those are physicians, nurse practitioners, or disability adjudicators) are called 'clinicians'. This is the language used through the dissertation.

2.3.1 Clinical Timelines

There are many tasks in the clinical domain for which (non-narrative) temporal language analysis is already known to be useful. The most obvious downstream application is the analysis of clinical records. Clinical records are produced at a large scale, and clinicians (even clinical researchers) have significant constraints on their time. To produce research on existing medical records, it often falls on automated ML methods to process text corpora. The prior section of this work showed an example of timeline extraction from within a single clinical text. This type of knowledge, which situates illness or level of ability in time, is vital for **disability adjudication**, in which a claimant's level of function *and the time span for which their function remains at that level* is compared against federal standards for work requirements.

This task forms the backbone of the work I contributed to on the NIH Epidemiology & Biostatistics research section (also called EpiBio) over the course of my doctorate. The NIH EpiBio team worked to classify and explore elements of patient disability. Though there are many competing definitions of disability (Pfeiffer, 1999), the EpiBio team builds its philosophy from the WHO ICF framework (WHO, 2001) (see Figure 17), which conceptualizes disability as an interaction between inherent internal factors (ex. congenital factors, acquired conditions) and external factors (home environment, work conditions, requirements of one's job). The social model of disability sees disability not as a binary state, but as a set of different levels across specific functions. Whether a patients' level of function is considered 'disabled' depends on their needs in their specific environment.

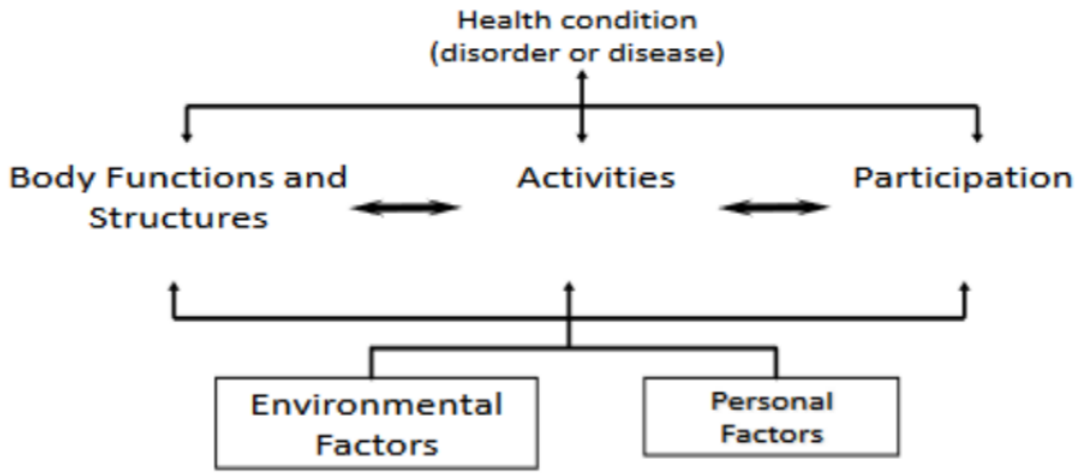

Figure 17: The ICF framework.

The NIH EpiBio group built models for extraction of function cues, segmentation of clinical text to build structural knowledge, and **clinical timeline generation** for disability adjudication. This last task is an area where two of this dissertation's disciplines (clinical text and temporal reasoning) intersect. The clinical timeline project and its challenges will be discussed in detail

in Chapter 4; here the work outlines the task itself and related frameworks.

A patient timeline collects and organizes all clinical records for an individual patient across a single timeline. Though this task is made easier with the introduction of electronic health records (EHRs), not all patients have a fully-digitized clinical history. For many illness types, significant printed medical documentation is required to make full sense of a patients' progression–when print records are transferred to digital, it is often difficult to identify record dates for sorting. An additional element present in some record types is *freeform testimonials* about patient experience. Though these testimonials are often given less authority compared to clinician observations, they provide irreplaceable information about the patient's day-to-day function. To that end, being able to extract an **ordered timeline** for all events in a single document helps to situate the record and reported function events in a single, aligned timeline.

A primary goal of EpiBio's clinical timeline work was to aid in disability adjudication for patient claimants. To meet governmental standards for disability, a claimant must express a consistent or acute level of function which prevents them from performing requirements of a job. Therefore, visualizing a patient's medical experience on the *temporal level* allows an expert adjudicator to focus on whether the level of function otherwise qualifies them for disability payments. Clinical timelines have a number of other potential downstream applications: understanding a patient's symptom progress in time can be critical for building prognosis, treatment plans, and even modeling illnesses and their progression across populations.

2.3.2 Patient-Clinician Communication:

> "Medicine is an art whose magic and creative ability have long been recognized as residing in the interpersonal aspects of patient-physician relationship."
>
> -*Communication of Affect Between Patient and Physician* (Hall et al., 1981)

The interpersonal aspects of clinical care are a significant topic within the field of medicine. In studies of these communications, researchers generally focus on the *form* of communication or the *content*. Here, 'form' is classified as the medium of communication rather than elements of grammar or rhetoric–those elements are instead grouped alongside the topics of conversation as 'content'.

Communication Content:

The practice of medicine is more than simply identifying illness and prescribing treatment. Interaction between clinician and patient shapes the process of practical medicine, in ways that can often be negative. According to one study on the subject, "Physician often talks jargon or seems not to heed patient's concerns. Mutual dissatisfaction is a frequent result" (Korsch et al., 1972). These communication issues are not easily solved; the work *Patient-Doctor Communication* notes that, "[Effective communication] is sometimes sacrificed with the intrusion of business into the patient–doctor relationship, the pressures of limited time for office visits, the culture of medicalization, and the sometimes all-consuming focus on technology" (Teutsch, 2003).

Further studies on patient-clinician communications are discussed in Chapter 3. Overall, these works show that communication is not separate to the healthcare process–it is a vital element of providing care. "Engaging patients and caregivers in the care process through sharing information, inviting their opinion, and collaborating with them constitutes another facilitator of patient-centered care and communication," states one study, "When patients and caregivers are engaged in the care process, misunderstandings and misconceptions are minimized. When

information is shared, patients and caregivers learn more about their health conditions and the care needed" (Kwame et al., 2021).

These recommendations seem to hold across cultures and medical communities[17]. Though cultural differences between patient and provider are noted as a potential barrier to successful communication, the need for successful communication appears to be true no matter who is seeking medical care. Therefore, the state of current communication is vital–the dissertation contributes a small glimpse into this ongoing concern among clinicians about their patient's insights.

Communication Form:

The field of healthcare communication has a number of competing priorities; many feel that monitoring the language clinicians use is less significant compared to the development of faster ways for patients and clinicians to communicate. Electronic health records (or EHRs) are a type of health records accessible to patients through online portals. As a new technology, they promised a more responsive type of patient-clinician communication, where patients could take more agency over their own healthcare journeys. This is standard practice in modern healthcare, but in the early days of EHR there was significant friction associated with their adoption. Studies on EHR adoption noted a trend where clinicians were more hesitant to adopt than patients. One concern of this issue was that EHRs would place additional demands on clinician time. Another was a persistent belief that the new visibility of EHRs would force clinicians to change the language of their records to avoid causing offense to patients (Ross et al., 2005[18]).

Language also affects clinical care in the area of treatment adherence. It is, naturally, a concern of clinicians that patients properly adhere to prescribed treatments. There are many studies examining factors in rates of adherence and nonadherence among patients; some show communication does play a role. Donovan et al. (1992) note that patients were more likely to adhere to treatment plans when given full explanations of that treatment and its side effects. Another noteworthy insight from Donovan's work is that patients showcased a need for *agency* during turbulent times of illness. Many in the study described their decision to modify or drop their prescribed treatment as one that allowed them to feel 'in control' of their healthcare journey–even when they understood how nonadherence would limit recovery (Donovan et al., 1992). This emotional need overrode other concerns in the patients' mind; the work will return to this issue of agency in Chapter 6.

2.3.3 Narrative Illness

This section examines the intersection between the *narrative* and *clinical* perspectives: the field of **illness narrative study**, which defines a genre of text where illness is described through narrative

An illness narrative is a self-account of an individual's personal experiences with illness, disability, or other chronic medical conditions. As a field, narrative illness study traces its roots to the book *The Wounded Storyteller: Body, Illness, and Ethics*, first published in 1995[19]. This book posits as its central thesis that illness self-reports function, fundamentally, as form of narrative (Frank, 2013). Frank's follows trends in a collection of illness self-reports, engaging in deep

[17] Chapter 3 cites studies from the United States, Iran, Ghana, and another diverse review of studies across sub-Saharan Africa.

[18] More references in Chapter 3.

[19] This dissertation cites the second edition from 2013.

readings to extract themes, pacing, and purpose from each text. It remains an influential work in the field to this day, especially for its two central theses: first, that stories are the means by which storytellers make sense of their world; second, that individuals dealing with the chaos of chronic and acute illness have a distinct need for such sense-making.

> "Selves are perpetually recreated in stories. Stories do not simply describe the self; they are the self's medium of being. [...] A first topic of this book is the need of ill people to tell their stories, in order to construct new maps and new perceptions of their relationship to the world."
>
> -*The Wounded Storyteller* and Second Edition Preface (Frank, 2013)

The act of *describing* the experience of illness, in Frank's view, is a method of self-reflection which is fundamentally transformational: to understand the experience of illness, he states, researchers must understand this fact.

Another significant work, *Narrative Medicine: Honoring the Stories of Illness*, was penned by a clinical practitioner (Charon, 2008). Where Frank emphasizes narrative as **self-reflection**, Charon focuses on the **communicative potential** of narrative. Unlike others in the field, her work does not identify narrative patterns in this genre. Rather, Charon's concern is to define an ideal relationship between clinician and patient; she urges clinicians to treat patients in a more humanistic manner and understand how narrative helps patients to explain their emotional needs. Charon theorizes that patients tell stories, even to clinicians, because it is only through stories that their entire personhood can be properly captured.

> "[Patients] enter whole [...] into sickness and healing, and their efforts to get better [...] cannot be fragmented away from the deepest parts of their lives. In part, this wholeness is reflected by—if not produced by the simple and complicated stories they tell each other. [...] Without narrative acts, the patient cannot convey to anyone else what he or she is going through."
>
> -*Narrative Medicine* (Charon, 2008)

Critical to the work in this dissertation are pre-existing typologies of illness narrative, where distinct patterns of behavior are examined through patient testimonials. The dissertation explores prior typologies here as a foundation for later work:

Frank's Typology:

Arthur Frank, in *The Wounded Storyteller*, defines the following typology of illness narrative (Frank, 2013):

1. **Restitution narrative**: the sequence of illness and recovery, told as a means of gaining self-confidence in one's own path to recovery.
2. **Quest narrative**: a story centered on an individual's actions during their illness, used to reclaim an internal sense of agency.
3. **Chaos narrative**: a complete breakdown of chronology; occurs in response to significant mental devastation.

Note that Frank's typology directly links *authorial intent* to testimonial content. In this dissertation's framework of textual construction, intent is an 'internal' factor under the control of a text's author. However, the **chaos narrative** is unique in that it is caused by an 'external'

source[20].

Hydén's Typology:

Another illness narrative typology comes from Lars-Christer Hydén. This typology frames illness narrative explicitly by *function* and *authorial intent* (Hydén, 1997). Hydén builds a more expansive understanding of illness narrative and how it may be used to process illness experience than Frank, defining the following narrative subtypes:

1. **Narrative construction of an illness world**: articulates and makes sense of the experience of illness.
2. **Narrative reconstruction of life history**: re-integrates illness experiences with one's sense of self.
3. **Narrative explanation and understanding of illness**: builds a cultural and philosophical understanding of the experience of illness
4. **Narrative as strategic device**: here, narrative serves as a tool to achieve some other goal in social interaction.
5. **Transforming individual experience into collective experience**: builds an emotional connection with a like-minded community, using the experience of illness narrative.

Bury's Typology

Finally, the work examines the typology proposed by Mike Bury. Bury's typology (Bury, 2001) covers the following narrative types:

1. **Contingent narrative**: a straightforward retelling of illness events. Bury proposes the contingent narrative is used for sense-making, but also may arise naturally through the need to teach others about an illness.
2. **Moral narrative**: a version of illness narrative which emphasizes the moral character of the author. In particular, the goal of a moral narrative is to emphasize to an audience that the author did not deserve their own illness.
3. **Core narrative**: the main goal of the goal narrative is self-affirmation. While the narrative might be shared with an audience, the primary benefit is for the author themself.

Returning to the sample text, it is clear this text fits neatly within Frank's definition of a **restitution narrative**: it straightforwardly discusses illness and recovery, and links recovery with a sense of confidence. It also fits Bury's **contingent narrative**, which is linked to the education of others about illness. Hydén's typology presents the most ambiguity for this particular sample of text–in some ways, it matches the **construction of illness world**, but the integration with the sense of self suggests it may be considered **reconstruction of life history**. Additional examination of the text from a clinical perspective continues in the next section.

2.3.4 Sample Application

From a clinical perspective, this text (Figure 18) describes a recovery path for an individual diagnosed with diabetes. Though this is a lifelong condition, the author has reached a point where treatment is preventing symptoms. Symptoms and function events can be extracted

[20]Though the mood connected to the author's illness journey originates within their mind, it is outside of the author's direct control.

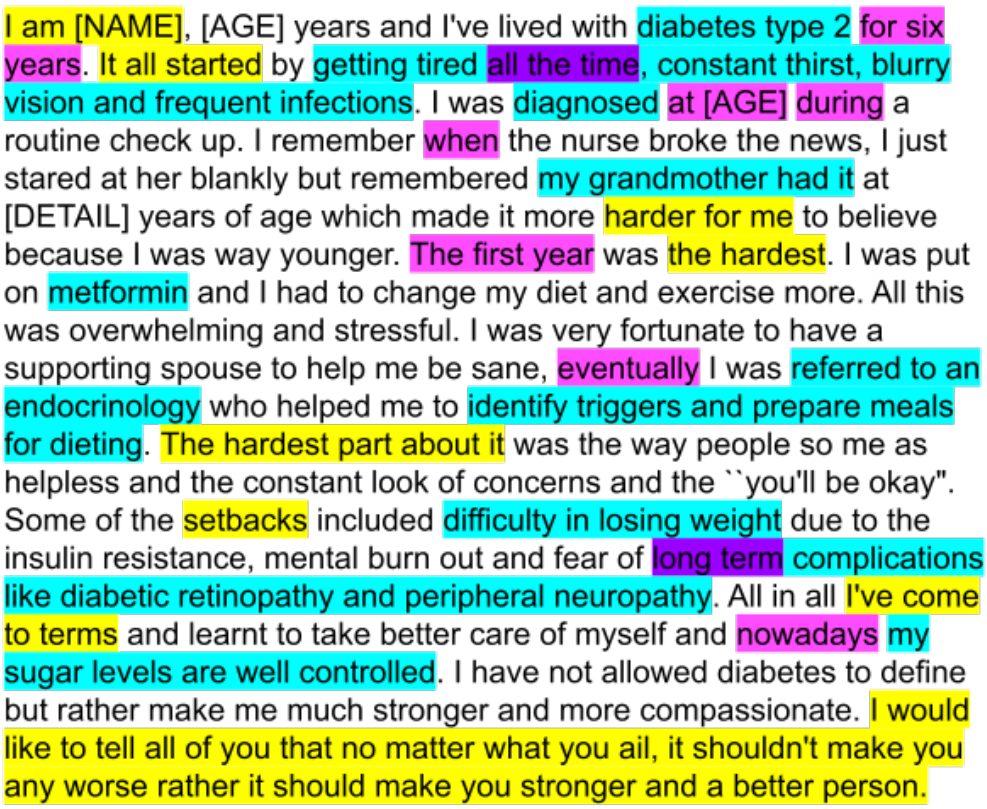

Figure 18: Sample text, annotated for all layers (time in pink, narrative in yellow, clinical in blue). Some phrases encode information for two layers.

(i.e. "getting tired all the time", "frequent infections", "difficulty losing weight"), potential long-term complications, and treatment information ("metformin", "identify triggers"). This clinical information could be used to develop a timeline of diabetes progression, which could then be compared against a cohort to draw conclusions about this author's particular experience and expectations. When researchers examines text for its clinical insight, it is not just medical 'jargon' that is captured, but all mentions which relate to the author's clinical experience and day-to-day function.

The clinical elements of this text can be described as follows:

1. **Pre-diagnosis**: Though the list of symptoms discussed in pre-diagnosis reflects a *narratively* short span of time, it can be inferred the symptoms caused enough interference in the participant's daily life to be noticeable.
2. **Diagnosis**: Many patients do not meet with clinicians as soon as they notice symptoms of chronic conditions. Whether due to economic constraints, clinician availability, or lack of knowledge, this participant only discovered their condition at a routine check-up. The diagnosis section also includes family medical history, which covers events much further 'past' than others on the narrative timeline.

3. **The first year**: The first year of treatment references early, difficult efforts to manage the condition. Medicine was prescribed along with diet and exercise changes; the fact that this narrative section is called the 'hardest' suggests these treatment options were not effective for managing the participant's diabetes.

4. **'Eventual' referral**: The referral represents a narrative turning point, but also references setbacks. The positives after referral were the identification of triggers and better dieting plans. Negatives focus on possible future events (i.e. long-term complications which did not occur but which worried the participant). This provides insight on patient priorities, and therefore what clinicians may wish to discuss with patients alongside treatment.

5. **Current day**: Finally, the present moment intersects with clinical elements of the narrative by saying that sugar levels are currently well-managed. This suggests the treatments after the turning point of endocrinology were the most successful for the participant.

Though this represents one anecdotal case study, applying this type of evaluation to cohort studies could help to surface treatments and resources which are most effective in managing a given condition. If multiple testimonies report only being diagnosed by chance during routine check-ups, it could support a movement to raise public awareness of diabetes symptoms. If the trend of symptoms improving after endocrinologist referral is common, researchers could argue a public health benefit to speeding up referrals for Type 2 diabetes patients.

2.4 Conclusion

The three frameworks discussed in this chapter explain distinct portions of the sample illness narrative testimonial. Areas of the text which feature few of one type of cue contain others, with the final output providing substantial coverage of the overall text. In this example, the initial framing of the piece clearly establishes time, and the conclusion is firmly rooted in a sense of narrative satisfaction, with clinical elements scattered throughout.

The core motivation of the dissertation's synergistic approach is that these three elements of text (and the three disciplines these elements draw from) build meaning through their interaction. Individually, these perspectives afford us insight into the text–but together, they synthesize an understanding that is more complete and comprehensive than any individual approach alone.

3 Related Works

This section of the dissertation discusses in more detail prior research from the three intersecting fields of the work. These related works are broadly organized into the topics "temporal reasoning", "narrative analysis", and "medicine".

3.1 Temporal Reasoning

Chapter 2 discussed the relationship between TEO/ETRE and other temporal reasoning tasks. Here, the dissertation expands this web (shown in Figure 19) and discusses related works in theses areas of NLP.

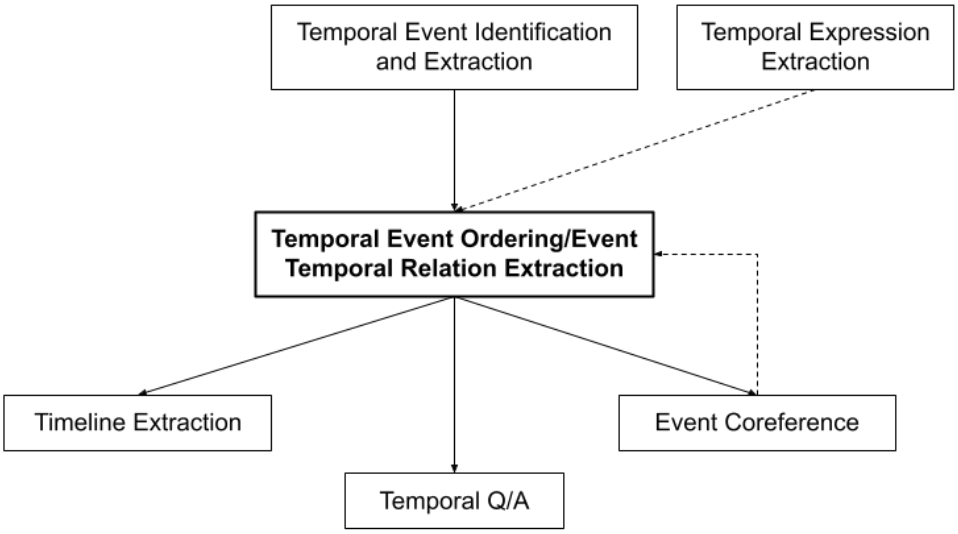

Figure 19: Web of NLP tasks which contribute to and which benefit from TEO/ETRE.

3.1.1 Timex Extraction

One new element of the expanded web is the Temporal Expression Extraction task, also known as **Timex Extraction**. This task is not strictly required for TEO/ETRE but provides long-distance contextual information for many models. A **timex** represents a (single- or multi-word) phrase within text that directly communicates temporal properties. As an example, the following phrases can all be considered as timexes, as they directly note a date:

- January 1, 2025
- 1/1/25
- The first of January '25
- The day after December 31, 2024

Though some versions of this timex are unlikely to appear in natural text, all should be comprehensible as corresponding to the absolute date **01-01-2025**. Some absolute time expressions rely on implicit knowledge about a document and context. One common example is the word "today"–placing it on a real timeline requires knowledge of the date of **document creation**[21]. The value of an absolute time expression for the task is significant: events linked to this time expression can be efficiently ordered relative to all other events with absolute time expressions (as shown in Figure 20).

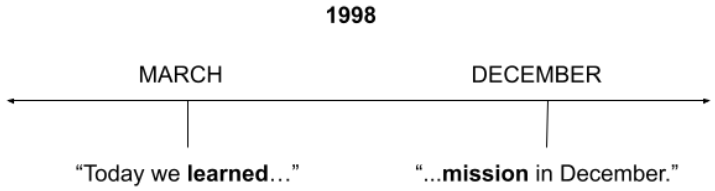

Figure 20: Given knowledge that document creation is 03/04/1998, events linked to absolute events can be ordered easily.

It is also possible for timexes to be relative to other events. Take an example from TimeBank[22]:

> "It wasn't until twenty years after the first astronauts were **chosen** that NASA finally **included** six women, and they were all scientists, not pilots."

The timex "[not] until twenty years after" establishes a duration of time between the event "chosen" and the start of event "included". This time expression is *relative* to another event, rather than *absolute*.

In the web of NLP tasks shown in Figure 19, Timex Extraction helps with–but is not necessary for–TEO/ETRE. TimeBank encodes both the absolute and relative types of timexes (the latter are formalized as 't-links') using manual annotation. There are limits to the TimeBank timex encoding[23], and not all elements of TimeBank timexes have been utilized by TEO/ETRE models trained on the corpus. Some models do integrate timexes as additional features, and see benefit from this inclusion (to be discussed in Section 3.1.3). However, manual annotation of timexes takes time beyond the already-challenging TEO/ETRE annotation task. When building new corpora, automation of Timex Extraction is often necessary to annotate at scale.

Timex Tools:

To properly extract a timex, a tool must do the following:

- Identify the boundaries of the timex. In the example timexes given earlier, it is necessary that the tool pull only the relevant phrasing from the main text.
- Convert the bounded timex into a standardized expression of temporal value. Multiple distinct phrases may refer to the same underlying date-time.

[21] In work based on the TimeBank corpus, the document creation annotation is available for all documents. But this is not necessarily true for other datasets which may be processed for TEO/ETRE.

[22] From document 'ABC19980304.1830.1636', an ABC transcript dated 03/04/1998.

[23] T-links contain the direction of event-event relations (i.e. before/after) but not additional duration information.

Many tools exist that perform this two-part extraction. **HeidelTime** (Strotgen et al., 2010) and **SUTime** (Chang et al., 2012) both use regular expressions to identify and match patterns for potentials timexes. These tools also convert matched regexes into standardized date-time values. There are inherent limitations to the regex method; Chang et al. (2012) note SUTime's inability to distinguish between homonyms[24] as one example.

The **Stanford CoreNLP** tool (Manning et al., 2014) improves upon this approach using a specialized parser within its Named Entity Recognition functionality. It leverages structural relations between words to extract time expressions from the surrounding text. Like HeidelTime and SUTime, this parser also returns a standardized time value. This same functionality has been exported by the Stanford team to the Python module **Stanza** (Qi et al., 2020). Stanza is easily incorporated into end-to-end Python pipelines and is used in the dissertation for some simple timex extraction.

It is necessary for a timex tool to identify the boundaries of a given time expression. However, the most common current tools ignore the grammatical language surrounding a time expression. This work argues that this short-distance context has significant impact on the meaning encoded in a time expression, and that omitting it limits the utility of timexes within TEO/ETRE. This dissertation proposes a new framework and tool for extracting time expression that is more sensitive to downstream work.

STAGE Framework:

The Semantic Temporal Alignment Grammatical Extraction tool (or **STAGE**) is a parser which extracts semantic details of time through grammar (Breitfeller et al., 2021). It produces similar but distinct results to Python packages such as Stanza, SUTime, and HeidelTime (Qi et al., 2020; Chang et al., 2012; Strotgen et al., 2010). STAGE is designed with a particular eye towards time as it is used in neural temporal models, encoding elements of grammar beyond word-level semantics.

When timexes are included in a text, they exist in relation to other objects or actions. An adverbial phrase most often modifies the clause's predicate, and time is often referenced within prepositional phrases which modify a noun. To demonstrate, consider the following parses, obtained with the Stanford CoreNLP tool (Manning et al., 2014)[25]:

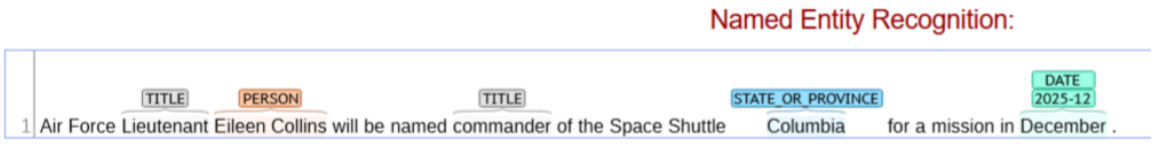

Figure 21: Named Entity Recognition recognizes 'December' as a timex.

In Figure 21, the word 'December' is recognized as a timex, and modern tools use the present date to infer that the reference is to the upcoming December in 2025[26]. However, there are grammatical elements to the temporal cue in this sentence, which can be more clearly seen through dependency parse (Figure 22).

[24]The word 'spring' can be a temporal season, or a verb relating to motion–but will always parse the same using regular expressions.

[25]Sample text also from TimeBank document 'ABC19980304.1830.1636', written in early 1998.

[26]An incorrect inference in this *particular case* because the sentence was written in March of 1998. But the inference follows correct logic given the grammatical style of the sentence *in isolation*.

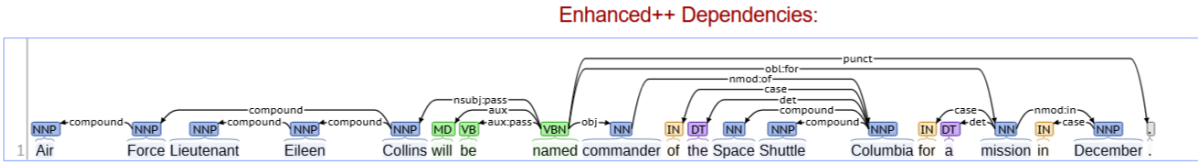

Figure 22: In the dependency parse, the word 'December' belongs to a prepositional phrase modifying 'mission'.

The timex does not exist in isolation; it modifies a specific event within that sentence. Suppose the sentence changes so that "in December" is the first phrase of the sentence rather than the last (see Figure 23). This changes the grammatical function of the phrase, which (in English) changes its semantic meaning. The sentence no longer cues a reader that a mission will take place in December–now, it is unambiguous that what occurs in that month is the act of naming Lieutenant Collins as commander.

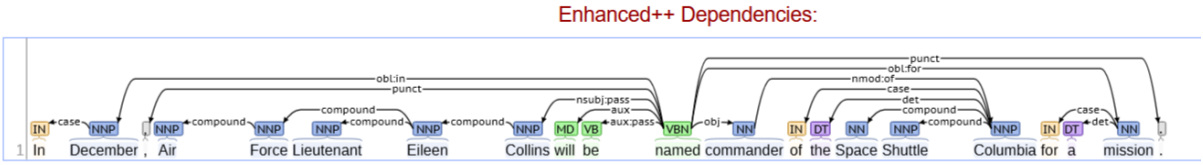

Figure 23: Now, the phrase 'December' belongs to modifies the predicate 'named'.

The semantic distinction between sentences whose timexes show different grammatical properties may seem trivial. But in practice timexes can be complex and require more careful processing. Consider a previously-referenced sentence from the same TimeBank document which contains a *relative* time expression:

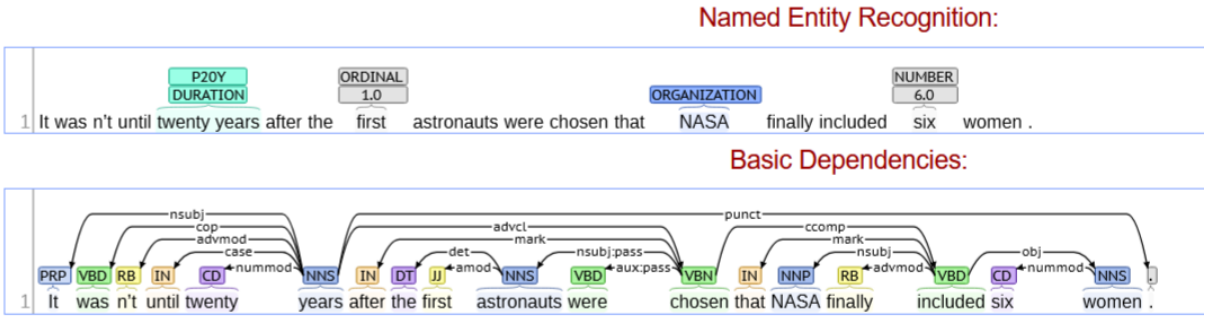

Figure 24: Two parses of the same sentence. The duration of 'twenty years' can be identified by NER, but its semantic meaning in the sentence is harder to extract.

In Figure 24, Stanford CoreNLP's NER can recognize the duration described by the phrase 'twenty years'. But those are not the correct boundaries of the time expression; no event within this text excerpt has a duration of twenty years, which is what NER identifies. It is clear to a human reader that this duration describes the time that passes between when the first astronauts were **chosen** and when women were **included**. The word choice used to

communicate this is structurally complex ("[not] until" uses negation to imply the starting point of the second event by explaining when this event *had yet* to start) and cannot be easily extracted even with the dependency parse.

The difficulties of this challenge case are met by the STAGE project. It is directed by two principles:

1. That the semantics associated with a timex change based on their grammatical properties within the sentence as a whole.
2. That the temporal semantics of timexes are better conceptually understood as *relative interactions* between timexes and one or more events in the text[27], rather than as isolated entities.

For example, in the sentence detailed in Figure 21, where CoreNLP identified the timex as "December", STAGE bounds the timex as "in December". In this case, there is no semantic distinction between the two. But for the example sentence in Figure 24, STAGE bounds the time expression as the full phrase "[not] until twenty years after". It works to identify not only the duration indicated by the temporal-lexical cues, but also that this cue marks a distance between its event and the noun phrase grammatically linked to the word "after". This output provides a richer understanding of the timex's semantic role in the sentence, one the dissertation argues is better suited for the temporal reasoning required for TEO/ETRE compared to competing frameworks.

3.1.2 Event Co-Reference

Event co-reference has a unique relationship with TEO/ETRE. Both are features of **event mention pairs** within a text. If the task central to the dissertation is to determine $Rel(E_1, E_2)$ for the label set assigned to Rel, co-reference determines $Ref(E_1, E_2)$, a boolean function which only asks if E_1 and E_2 are mentions which reference the same underlying event.

The question then becomes: at what point are two mentions co-referent? Quine (1985) notes the vagueness behind the concept of a differentiated 'event', but builds a possible definition for co-reference in which both mentioned events occupy the same space, exist during the same time interval, and involve the same objects. Davidson (1969) also discusses the temporal element, calling it necessary but not sufficient: "Of the temporal element of co-reference, I am uncertain both in the case of substances and in the case of events whether or not sameness of time and place is enough to insure identity" (Davidson, 1969).

Song et al. (2015) build on frameworks such as these to produce a more thorough definition of event-pair co-reference. Song proposes that co-referent events must feature the same *arguments*, which include:

1. Semantic agent and patient.
2. Spatial location.
3. Temporal interval.
4. Realis attribute (whether the event occurred in real time or is hypothetical).

This type of co-reference definition is used in certain challenge tasks–as documented in Song et al. (2016)–though others less rigid and more intuitive definitions of co-reference (Mitamura et al., 2017).

[27] This approach builds on the existing TimeBank t-link formalization, which are underutilized in TEO/ETRE.

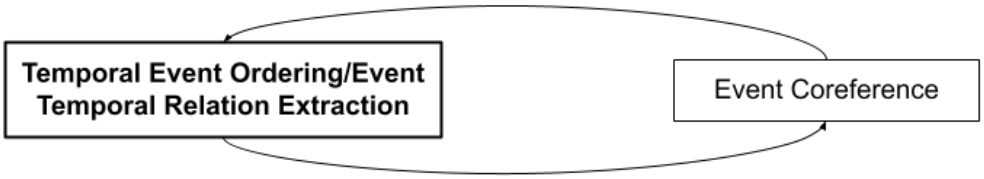

Figure 25: The relationship between TEO/ETRE and co-reference.

While co-reference can be downstream from TEO/ETRE, it is not strictly a downstream task. Like timex extraction, co-reference can also inform TEO/ETRE prediction (see Figure 25). Because *temporal sameness* is required by co-referent objects, $Set(Ref(E_1, E_2) = true) \subset Set(Rel(E_1, E_2) = simultaneous)$. It is possible to therefore make inferences about each attribute using the other:

1. If $Ref(E_1, E_2) = true$, then $Rel(E_1, E_2) = simultaneous$
2. If $Rel(E_1, E_2) \neq simultaneous$, then $Ref(E_1, E_2) = false$

In theory, NLP models predicting co-reference should synthesize well with TEO/ETRE work; each bolsters predictions for the other. SOTA co-reference models include Peng et al. (2016), Lu et al. (2017), and Zhao et al. (2025).

3.1.3 TEO/ETRE Models

This section describes past TEO/ETRE models across multiple categories.

Logical Temporal Models:

Here the dissertation will discuss both logical frameworks which underpin later temporal modeling efforts, and temporal models which rely on semantic logic.

Several papers by James Allen define and refine the logic behind temporal language. Allen (1983) presents the **interval framework** of time, while Allen (1984) formalizes temporal logic. Allen et al. (1985) and Allen (1991) returns to and expands the 13-label event-pair relation framework. Modeling work tends to abridge this framework in practice. The TEO/ETRE challenge task defined in UzZaman et al. (2013), for example, features a curated subset of temporal relations, as does the TimeBankDense dataset built and annotated specifically for TEO/ETRE (Cassidy et al., 2014). Other works explicitly return to Allen's logical frameworks, like Huang et al. (2023) which uses the interval perspective of time to better conceptualize event-pair ordering.

A second logical framework relevant for TEO/ETRE is **script theory**, presented by Schank et al. (1975) and refined for NLP by Chambers et al. (2009). Scripts build structural representations of regular relationships between events that would be known to humans within a shared cultural space. When multi-part events are frequent enough to be considered 'common knowledge', this may influence the order in which they are presented in a text by human authors. For example, in a culture where baking cakes is common, it is likely to be understood without explicit statement that a cake must be *baked* before it can be *frosted*. Therefore, a human text may be less concerned with using textual order to make that sequence clear. The impact of

script theory on how humans choose to order 'common-sense' events is highly contextual, and relies on external knowledge.

Finally, some NLP models utilize logical elements of time to scaffold statistical decision-making. Bramsen et al. (2006) enforce logical frameworks of temporal transitivity on more traditional NLP features. This formalizes the **transitivity entailments** used throughout this dissertation. Other temporal models of time with logical elements include Mani et al. (2006), Lapata et al. (2006), UzZaman et al. (2010), Llorens et al. (2010), and Reimers et al. (2018).

Statistical Temporal Models:

Chapter 2 briefly discussed past models for the task of TEO/ETRE. Here, performance of these models is showcased against standard benchmarks: TimeBankDense (Cassidy et al., 2014), TD-Discourse (Naik et al., 2019), and MATRES (Ning et al., 2018). Each of these benchmarks uses text and some associated annotations from Pustejovsky et al. (2003b)'s TimeBank news article corpus. Existing SOTA for this task can be roughly divided into three categories:

1. **Direct Text Encoder Models**: Ning et al. (2017); Cheng et al. (2017); Han et al. (2019b); Han et al. (2019a); Ballesteros et al. (2020).
2. **Structural/Graph Models**: Beltagy et al. (2020); Kitaev et al. (2020); Zaheer et al. (2020); Liu et al. (2021); Mathur et al. (2021); J. Zhou et al. (2022).
3. **Recent Miscellaneous SOTA**: Man et al. (2022); Huang et al. (2023).

Short-distance models typically build off a text encoder, collecting the **context window** of text which immediately surrounds events in a pair[28]. Ning et al. (2017)'s SP+ILP model added inference to a structured perceptron model while Cheng et al. (2017) presented a traditional BiLSTM approach. The later Han et al. (2019b) introduced MAP capability to produce BiLSTM + MAP, and Han et al. (2019a) leveraged deep learning for the same problem (SSVM). Finally, Ballesteros et al. (2020) demonstrated an early multi-task model which still emphasized short-distance capture. These models performed best on the TimeBankDense and MATRES models but struggle on TDDiscourse-Man, which contains the highest proportion of long-distance event-pairs. Performance for all baselines can be found in Table 1.

	MATRES	**TB-Dense**	**TDD-Auto**	**TDD-Man**
SP + ILP	76.3	58.4	46.1	23.8
BiLSTM	59.5	48.4	51.8	24.3
BiLSTM + MAP	75.5	64.5	57.1	41.1
DeepSSVM	-	63.2	58.8	41.0

Table 1: Performance of transformer-based models on TEO/ETRE.

When benchmarks with long-distance pairs became more prominent within the field, models began to incorporate structural and long-form information. Longformer (Beltagy et al., 2020), Reformer (Kitaev et al., 2020), and BigBird (Zaheer et al., 2020) encoded long-distance structural information with distinct attention mechanisms[29]. Another key method for encoding long-distance structural information was the inclusion of Graphical Neural Networks (GNNs) to replicate the syntactic, semantic, and temporal understanding of a human reader. Liu et al.

[28]In the Temporal Reasoning section of this chapter, *context* refers to the segment of text given as input to these encoders.

[29]The specific implementation of these models in this work's tables follows the implementation from Man et al. (2022).

(2021) introduced UCGraph with a GNN component, while Mathur et al. (2021) used three GNNs in TIMERS pipeline. One of these GNNs incorporated timex information from TimeBank annotation (but not t-links). J. Zhou et al. (2022) introduced relational graph elements to build RSGT. Baseline performance can be found in Table 2.

	MATRES	**TB-Dense**	**TDD-Auto**	**TDD-Man**
Reformer	-	-	65.9	43.7
BigBird	-	-	65.3	43.3
UCGraph	-	59.1	61.2	43.4
TIMERS	82.3	67.8	71.1	45.5
RGST	82.2	68.7	-	-

Table 2: Performance of long-distance models on TEO/ETRE.

The introduction of long-distance structural elements generally improves performance for all categories, though the most notable improvement was in TDDiscourse-Auto and TDDiscourse-Man (the datasets focused on long-distance event-pairs).

Most recent SOTA models tend to produce specific innovations on the structural/graph approach which do not fit into neat categories. SCS-EERE (Man et al., 2022), rather than altering existing model architecture, used an algorithmic selection process on data to select vital context sentences that improve performance during training. Huang et al. (2023) introduced the Unified-Framework (UF) which returned to Allen (1983)'s interval logic, conceptualizing events as start- and endpoints and using predictions for the relations between all start/endpoints in a pair to scaffold traditional TEO/ETRE.

	MATRES	**TB-Dense**	**TDD-Auto**	**TDD-Man**
UF	82.6	68.1	-	-
SCS-EERE	83.4	-	76.7	51.1

Table 3: Performance of other SOTA models on TEO/ETRE.

Note that certain models were not run on specific datasets in prior work, and thus lack scores to compare against here. Overall, encoder models perform best on the TimeBankDense and MATRES models but struggle on TDDiscourse-Man, which contains the highest proportion of long-distance event-pairs. The introduction of long-distance structural or graph elements generally improves performance for all categories, but show particular improvement on TDD-Auto and TDD-Man.

3.1.4 Large Language Models

Large Language Models (LLMs) currently dominate the NLP sphere. These new models bring with them great promise for solving problems of natural language generation, producing human-like text and speech for a wide range of user prompts. However, these models do possess limitations, explored through various benchmarks within recent years. In particular, LLMs showcase certain difficulties when it comes to tasks involving **temporal reasoning**.

In theory, any current LLM could be utilized for the task of TEO/ETRE if given the correct prompt. S. Chen et al. (2025) tests LLMs against not only objective but *subjective* perceptions of time, in an interesting expansion of the task. Other salient uses of LLMs in the temporal

sphere include data generation; a recent pre-print study has produced a clinical dataset annotated by a version of LLaMA for temporal properties (H. Wang et al., 2020), and asserts that this annotation data can fine-tune existing BERT models for downstream QA with 10% improvement. "A GPT-2 model fine-tuned on our dataset generates more clinically relevant responses" it claims[30].

These works represent potential for LLMs in the field of temporal reasoning, at least as far as temporal properties inform downstream tasks. Next, the work explores specific LLM benchmarks for time.

LLM Temporal Benchmarks:

> "The complex and diverse associations between temporal information and entities in knowledge-intensive scenarios substantially hinder models' ability to accurately correlate time with facts."
>
> -*TIME: A Multi-Level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios*, Wei et al. (2025)

Though the flexibility of LLMs compared to other models is immense, there remain concerns that the general LLM may under-perform on specific tasks. In studies of LLMs on temporal reasoning, there is a consistent trend of these models struggling for predict for complex reasoning tasks.

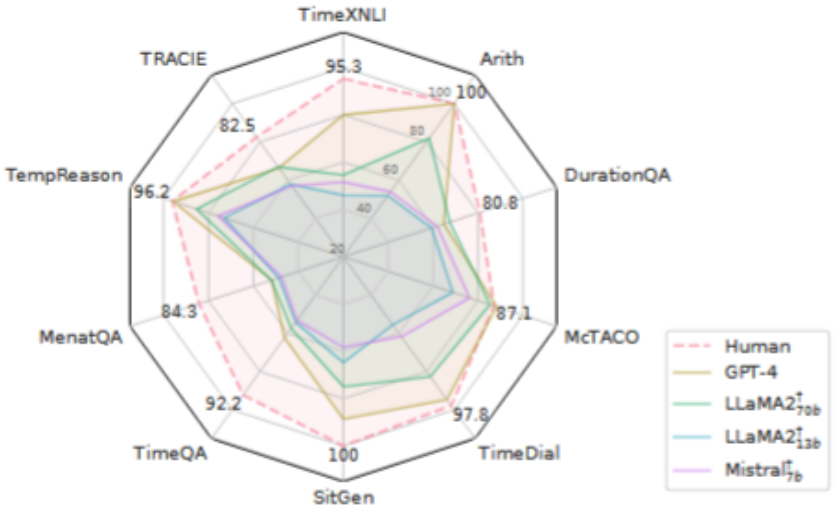

Figure 26: Performance of SOTA LLMs on TimeBench temporal reasoning tasks, from paper (Chu et al., 2023). Note the comparative performance on TRACIE, MenatQA, and TimeQA versus human baseline.

Chu et al.'s **TimeBench** benchmark study explored the performance of multiple SOTA LLMs on distinct temporal NLP tasks and compared their performance against a human baseline (Chu et al., 2023). The team highlighted the performance of LLMs on the TRACIE (temporal

[30]It must be acknowledged that this statement is qualitative. The paper does *not* discuss the criteria it used to define clinical relevance.

ordering between implied events in the text), MenatQA (counter-factuals and time-disruptive questions to induce implicit reasoning), and TimeQA (QA of time-sensitive topics with explicit and implicit time-involved reasoning) benchmark tasks as compared to human baseline (see Figure 26).

Though GPT-4 was noted as a high performer among the LLMs, even it fell drastically below human performance. The study attributes this failure to a limitation in LLMs ability to perform *multi-hop* and *implicit* symbolic temporal reasoning. When models had to extract contextual information from the text to complete a task, the more LLMs struggled. Overall, the study concluded that LLMs did not achieve the requirements of such temporal tasks: "Our findings indicate a substantial gap between state-of-the-art LLMs and human performance, emphasizing the need for further research in this area." (Chu et al., 2023)

The **TIME** benchmark (Wei et al., 2025) does not compare LLMs to human performance, but the associated paper uses LLM accuracy scores as a means to highlight current gaps in LLM temporal prediction. Wei et al. note Order Reasoning (temporal reasoning with ordinal expressions), Relative Reasoning (interpretation of relative temporal expressions), Co-temporality (understanding of events which overlap in time), and Timeline (relative ordering of event triples)[31] as areas where even high-performing models struggled to predict for gold-standard data (see Figure 27).

	Level-2			Level-3		
	ER.	**OR.**	**RR.**	**Co-tmp.**	**TL.**	**CTF.**
Deepseek-V3	75.69	39.77	41.76	46.62	10.00	44.82
Deepseek-R1	78.20	**57.09**	**57.79**	47.45	**33.33**	**55.71**
GPT-4o	80.68	45.83	46.56	45.45	20.00	50.72
OpenAI o3-mini	**82.24**	52.62	48.98	**54.34**	**33.33**	52.07

Figure 27: Abridged table from TIME paper (Wei et al., 2025), with emphasis on model accuracy for OR, RR, Co-Tmp, and TL tasks.

Like TimeBench, they reference the challenge *implicit* knowledge poses to LLMs: "Knowledge intensive events makes it challenging for capturing complex temporal expression and relationship [...] models face significant challenges in comprehending implicit temporal expressions and intrinsic temporal relationships between events. This phenomenon suggests that the complex and diverse associations between temporal information and entities in knowledge-intensive scenarios substantially hinder models' ability to accurately correlate time with facts." (Wei et al., 2025).

The **TRAM** benchmark by Y. Wang et al. (2023) is, overall, more optimistic about the potential for LLMs, particularly with the use of alignment techniques like multi-shot prompting and Chain-of-Thought (CoT) which they found improved performance across tasks. However, they note a specific case for the temporal Relation tasks, a prediction task that forms the center of this dissertation . In temporal Relation prediction, even small-scale BERT and RoBERTA-based models significantly outperformed compared to LLMs[32] (see Figure 28). This leads to

[31] Similar to the TEO/ETRE task which we will focus on in this dissertation.

[32] With best performance using reStructured Pre-training model RST (Yuan et al., 2022) which predates widespread LLMs.

Model	Rel. *Acc./F1*
Random	33.3/33.3
Llama2 (0S, SP)	35.2/33.1
Llama2 (0S, CoT)	40.1/38.5
Llama2 (5S, SP)	38.1/36.6
Llama2 (5S, CoT)	43.0/41.3
Gemini (0S, SP)	60.5/60.1
Gemini (0S, CoT)	64.2/63.6
Gemini (5S, SP)	62.8/62.3
Gemini (5S, CoT)	65.1/64.9
GPT-3.5 (0S, SP)	40.5/39.1
GPT-3.5 (0S, CoT)	44.1/42.9
GPT-3.5 (5S, SP)	42.5/41.3
GPT-3.5 (5S, CoT)	45.9/45.2
GPT-4 (0S, SP)	60.6/58.8
GPT-4 (0S, CoT)	63.6/62.9
GPT-4 (5S, SP)	62.0/61.5
GPT-4 (5S, CoT)	66.5/65.2
BERT-base	86.5/86.6
BERT-large	89.5/89.5
RoBERTa-base	87.0/86.8
RoBERTa-large	90.0/90.0
RST	**91.5/91.6**
Human Experts	96.0/96.0

Figure 28: Abridged table from TRAM paper (Y. Wang et al., 2023). Note the performance of the BERT suite of models compared even to GPT-4o with alignment techniques.

a critical conclusion, regarding these models and their relative performance: "**Sheer model size does not always equate to superior performance.**" (Y. Wang et al., 2023).

Other LLM benchmarks of note, both in the field of temporal reasoning and outside it, include the following:

- **TempUN**, Beniwal et al. (2024): The intention behind TempUN was to test LLMs' abilities to understand knowledge which changes over time. This is a related task to TEO/ETRE prediction[33]. Beniwal concludes that "LLMs lacks temporal reasoning and understanding capabilities," though "open-source LLMs perform better than closed-source models on the average scores of all six [multiple choice question]-based evaluations."
- **TempBench**, S. Chen et al. (2025): TempBench poses psychological questions about the human experience of time (where the same objective duration may feel longer or shorter). It reported "a substantial gap between the performance of LLMs and that of humans" in this task.
- **SocialIQa** and **ToMi**, Sap et al. (2022): Two benchmarks for social and emotional reasoning (SocialIQa) and understanding of external mental states (ToMi). Sap noted

[33]Understanding the relative order of events requires understanding when a state event changes or 'ends'.

generative models fell consistently below human and specifically warned that "Increasing model size might not be enough to reach human-level accuracy" in the future.

- **Multi-Axis Logical Evaluation Framework** Xu et al. (2025): A framework for evaluating logical reasoning along six axes (Correct, Rigorous, Self-aware, Active, Oriented, and No hallucination). The study showed that LLMs often fail at orienting to the correct direction of logical reasoning.

Though LLM alignment methods such as multi-shot prompting and CoT can fill some gaps in LLM reasoning, there is skepticism that this methodology will be useful in all cases. Y. Wang et al. (2023) showed baseline LLMs could improve performance on most aspects of TRAM using 5S and CoT settings, but Chu et al. (2023) reported that "introducing zero-shot CoT prompting results in consistent declines [...] in the few-shot scenario, CoT prompting also fails to yield consistent improvements, varying depending on the task." CoT did seem most useful to improve symbolic and multi-step reasoning, two areas shown to be necessary for complex temporal reasoning tasks (Chu et al., 2023), but as Y. Wang et al. (2023) noted, even that improvement did not allow LLMs to exceed the performance of smaller, targeted BERT/RoBERTa models.

Other Concerns:

An additional, growing concern with the use of LLMs is the issue of *interpretability*. As Anthropic CEO Dario Amodei noted in 2025:

> "People outside the field are often surprised and alarmed to learn that we do not understand how our own AI creations work. They are right to be concerned [...] Looking inside these systems, what we see are vast matrices of billions of numbers. These are somehow computing important cognitive tasks, but exactly how they do so isn't obvious."
>
> -*The Urgency of Interpretability*, Amodei (2025)

The 'black box' problem is not new to LLMs. However, the scale of these large models further exacerbates a common issue within NLP. While tools exist to explain the behavior of smaller, more focused neural models, the field falls behind in explaining how LLMs come to their decisions. There is an obvious benefit to clearer and more explainable models within this task space, especially when (as has been noted) "sheer model size does not always equate to superior performance" (Y. Wang et al., 2023).

3.2 Narrative Analysis

Here the dissertation discusses work within the broader field of narrative analysis.

3.2.1 Time in Narrative

A foundational assertion of the dissertation work is that non-chronological orderings of time in text are less comprehensible to a human reader–therefore, deviations from chronological must have some unconscious motivation that makes up for this loss of clarity. Zwaan (1996) examines reading time (i.e. how long it takes for humans to process a written document) for texts containing distinct representational relations, finding that reading time increases when subsequent events have significant gaps separating their *real-time* occurrences[34]. Kelter et al.

[34]Though TEO/ETRE does not care about the duration of time between events (only their order), long gaps increase the possibility that some temporal deviations will occur at some point in the text.

(2004) notes that reading comprehension decreases when asked about events which took place long ago in real-time even when they were mentioned recently in a narrative–once again, the position of an event mention *in time* is a distinct property for human readers, and divergences between this position and their position *in text* do impact human readers.

3.2.2 Narrative in Computation

Though "narrativity" and elements of text associated with it tend to be more broadly defined than is typical for NLP tasks, there are many existing studies which have explored these elements of text through a machine-learning lens. This section discusses models for narrativity, semantic language sub-types, rhetorical structure, and even definitions of authorial intent. Piper et al. (2021) makes an effort to quantify "narrative" itself within large-scale text analysis. Follow-up work had human annotators label texts which described occurrences of events for ratings of "narrativity" and then performed large-scale text analysis to identify associations among text by label (Piper et al., 2022). They found three types of language which were most likely to cue human readers to perceive a text as more narrative-like:

1. Experientiality: the state of having an agent who is experiencing the events in the text.
2. World-making: the situating of events in text in a setting, whether that be the real world or a fantastical location.
3. Sequentiality: the situating of events in text within time.

This definition of narrative has been used in later studies: Bamman et al. (2025) examined the level of "narrative" language in song lyrics using the above attributes as proxy, as one example.

3.2.3 Computational Linguistics

This dissertation uses specific tools from the field of computational linguistics. One is LIWC-22, a collection of specific dictionaries which sort words into one or more categories (ex. linguistic, rhetorical, figurative) to track psychological and social trends within text (Boyd et al., 2022). LIWC dictionaries allow a more focused analysis of large documents compared to raw word counts, and in particular provide insight for discriminant analysis.

There are tools which extract rhetorical elements of text like discourse acts. Rhetorical parsers build a tree structure of rhetorical elements within an input sentence, to better understand the specific contributions each element makes to the final message communicated by the text. It is therefore related to authorial intent, the focus of this dissertation. Integration of rhetorical structure can improve performance on lexical NLP models (Bhatia et al., 2015)–one SOTA model the work builds on uses discourse parsers by Ji et al. (2014) and the 'discoursegraphs' project by Neumann (2015).

Multiple NLP works define taxonomies of authorial intent. Kruk et al. (2019) builds a model seeking to predict the intended purpose of posts on social media and defines 8 classes of "rhetorical intent". Abu-Jbara et al. (2013) is a work in bibliometrics which identifies how and why papers are cited. Its taxonomy features 6 categories of intent. There is little arguable overlap between these two taxonomies; in both cases, the taxonomy is heavily specialized to its associated domain and downstream task. Authorial intent remains a difficult property to quantify and especially to generalize across domains.

A recent work uses LLMs to examine intention across text: an experiment by Dutt et al. (2024) prompted LLMs to parse conversations for human intention information (called INT), with

notable success. The INT values from this study were produced in a freeform format, with significant variety among labels. This means that these labels are not necessarily suited to strict classification of documents based on broader narrative intention category, as is done in this dissertation. However, they represent potential within the field for ML-focused analysis of authorial intent.

3.3 Medicine

This section discusses studies of medicine (both the language around medicine and medical studies) which inform the dissertation research questions.

3.3.1 Language in Clinical Text

The language of clinical text is key to several downstream tasks of clinical analysis. This section discusses language around disability, patient-clinician communications, and how the framing within communication impacts other elements of treatment.

Disability Adjudication:

Definitions of 'disability' vary among works. Pfeiffer (1999) noted a wide variety of existing definitions, some of which hinge on chronic inability to work and others which emphasize some limitation to some element of life. Fried et al. (2004) discussed a relationship between comorbidity, frailty, and the attribute of disability. Leonardi et al. (2006) noted a state of 'objective' disability, the description for which may differ from an individual patient's satisfaction with their level of function.

In modern clinical settings, many follow the **social model of disability**, where distinct types of patient function interact with one's home environment and personal responsibilities. This definition is the foundation of the WHO's ICF framework (WHO, 2001).

Patient-Clinician Communication:

> "I got quite upset to my doctor who then said 'Oh I didn't know you were like that' because I started to cry in the surgery, he said 'I always thought you were such a, you know, someone with common sense, I didn't know you got upset like this' and put me on tranquillisers."
>
> -Anonymous Patient from Wilson et al. (2006)

The impact of patient-clinician communication on a patient's emotional and physical well-being is significant, and this impact has been a concern of clinical researchers for decades. Though clinical practices change over time in response to new scientific insights, the problem of communication remains to be solved.

Korsch et al. (1972) noted a divide in register between patient and clinician during conversations, and blames this divide for friction in their professional relationship. Another study explore the impact of specific verbal word choice in face-to-face patient-physician communication, reporting a positive relation between the wording of the clinician and patient satisfaction (Hall et al., 1981). Ong et al. (1995) also attributed "dissatisfaction" among certain patients "in part [to] doctors' communicative behavior". Wilson et al. (2006) reported that, in patients with chronic conditions, they perceive better emotional communication from nurses than clinicians.

Teutsch's work reflected a more recent but standardized how-to on patient-clinician communication (Teutsch, 2003). However, more specific primers exist for individual communities.

In 2021, a literature review of existing communication studies confidently recommended a patient-centered paradigm to clinical communication called PC4 where trust leads to natural conversation where "the content of communication [...] is both 'personal' and 'explanatory'." (Kwame et al., 2021). "Engaging patients and caregivers in the care process through sharing information, inviting their opinion, and collaborating with them," Kwame stated, "constitutes another facilitator of patient-centered care and communication. When patients and caregivers are engaged in the care process, misunderstandings and misconceptions are minimized. When information is shared, patients and caregivers learn more about their health conditions and the care needed." Another review likewise recommended an approach which positions "provider as confidant" and allows "patients [to] be empowered in care processes" (Camara et al., 2020).

Donovan et al. (1992) studied treatment adherence, and how it intersects with the patient-clinician relationship, while Funnell et al. (2000) directly recommended a relationship focused on patient needs to reduce non-adherence. Though *language* is not a direct focus of treatment adherence studies, communication between patient and clinician necessarily requires it.

Communication Form

Electronic health records (or EHRs) are standard in modern healthcare, considered to improve communication between clinicians and to increase the accessibility of healthcare information to patients. Nonetheless, roll-outs of EHR tools were met with friction. Many studies examined the sources of that friction and provided theories on how to alleviate it (Hassol et al., 2004; Pyper et al., 2004; Ross et al., 2005; Delbanco et al., 2010; Delbanco et al., 2012; S. S. Woods et al., 2013; Jilka et al., 2015; Grünloh et al., 2016; Tieu et al., 2015). In general, studies reported that clinicians were more hesitant to put healthcare information into EHRs than patients were to access them.

3.3.2 Therapeutic Modality

Certain aspects of therapeutic study (without direct focus on language) are also relevant to this dissertation work. This section discusses the psychology of memory, medical perspectives on illness narrative, and other talk therapies.

Psychology of Memory and Time:

Human memory exists at an intersection between meaning and time. It is generally considered to be common knowledge that the human process of recalling memories requires *reconstruction* of temporal order. To quote Van der Kolk (1998): "At least since 1889 [...] it has been widely accepted that what is now called declarative or explicit memory is an active and constructive process." Tulving (2002) described this type of memory as **episodic memory**, memories which are rooted in events situated in time. This reconstructive process is not foolproof; distortions in the construction of episodic memory can change how these memories are communicated to others.

Factors that may influence human recall include:

- **General memory issues:** Cognitive psychology asserts that memory is not retained in chronological order and may be distorted on reconstruction (Friedman, 1993). This issue is exacerbated for experiences of illness, which can further damage memory in the cases of mental conditions or overall stress. Studies have shown patients with Alzheimer's, for example, produce "less chronologically organized" narratives than a control group event when asked to relay events in short-term memory (Usita et al., 1998).

- **Trauma:** Trauma is known to affect the storage of memories related to that trauma, leading to either extreme of "retention" (where the traumatic memory is clearer than non-traumatic memory) and "forgetting" (where the memory is distorted to a higher degree than is normal) (Van der Kolk, 1994; Van der Kolk, 1998). In both cases, the effects of trauma on memory carry implications for analysis of illness narrative, in which participants' experiences may qualify as traumatic.
- **Cultural and language barriers:** Many genres of text behave differently depending on the native culture and language of those creating them. Work on illness narrative that focuses on native English speakers may only be applicable to native English speakers, and might not generalize outside that demographic.
- **Responsivity of patient to illness narrative format:** In a critique of narrativity theory, philosopher Strawson posits that individuals can fall into one of two types: *diachronic* and *episodic* narrators. Diachronic narrators, Strawson proposes, perceive of their own experiences as a continuous timeline and therefore respond better to attempts to "narrativize" them, while episodic narrators may simply not produce the same types of narratives expected by experts in the field (Strawson, 2004).

Illness Narrative and Criticism:

"Illness narrative" presents as a potentially-rich source of linguistic data. Work by William Labov on the "flow of speech" proposed that there are distinct subsets of human experience more likely to prompt unfiltered communication, of which "death" is one super-category (Labov, 2013). Illness, as a related topic, may likewise produce unfiltered disclosure. It is interesting to note that Labov associates the category of death with the emotive element of "thrill", while illness narrative (as it is formulated by Arthur Frank) tends to link the topic with a wide range of emotional responses.

It is true that illness narrative as a genre presents a variety of texts and text types. Despite this, critics of the field argue Frank's definition and perspective on illness narrative remains overly reductive in how it treats ill individuals. These critics assert there is a broader scope of human experiences in healthcare beyond those showcased in *The Wounded Storyteller* and similar works (A. Woods, 2014). Others feel that despite the field's history, it has yet to yield data with rigorous analytic impact; the typologies the field produces (detailed in Section 2.3.3) are as yet only backed by qualitative evidence rather than statistical significance (Atkinson, 2009).

Another concern is that the real-to-text transformation that occurs in illness narrative might obscure ground truth and therefore provide less useful information for researchers. This process of narrative self-reflection could also end up doing harm to patients during an already-chaotic period of their lives by reinforcing patterns of trauma (both critiques posed by Strawson, 2004). Theses responses are similar to critiques levied against talk therapies, discussed next.

Talk Therapies:

Many schools of psychology encourage 'debriefing' about traumatic events similar to illness narrative–though there are caveats. Consider the following modalities of talk therapy, which do appear to validate the foundational thesis of illness narrative: that revisiting a traumatic experience through narrative can be healing.

1. Cognitive Behavioral Therapy (CBT), as defined by the American Psychological Association, "focuses on the relationship among thoughts, feelings, and behaviors" (APA, 2017a). It is asserted in CBT that changing the way an individual thinks about prior experiences

changes their response to new experiences, and that this can be leveraged in positive ways to teach effective coping mechanisms for trauma. CBT is "strongly recommended" by the APA to deal with traumatic conditions like PTSD through guided discussion. "Exposure to the trauma narrative, as well as reminders of the trauma or emotions associated with the trauma, are often used to help the patient reduce avoidance and maladaptive associations with the trauma" (APA, 2017a).

2. Critical-Incident Stress Debriefing (CISD) is a treatment practice used for first responders, who are often exposed to sights in the course of their duty which may be emotionally affecting. Organizations like OSHA recommend debriefing soon after such incidents, asserting that this practice allows individuals to process these incidents and prevent them from becoming long-term stressors (OSHA, 2021).

3. Narrative Exposure Therapy (NET) is considered by the APA to be "conditionally recommended" for treatment of PTSD. It is used primarily in communal therapy settings. The APA's recommendations on it show a strong concern with temporal understanding. "With the guidance of the therapist, a patient establishes a chronological narrative of his or her life," it asserts, "concentrating mainly on their traumatic experiences, but also incorporating some positive events" (APA, 2017b). APA discussion of NET also highlights narrative as something which produces order from chaotic experience: "By expressing the narrative, the patient fills in details of fragmentary memories and develops a coherent autobiographical story. In so doing, the memory of a traumatic episode is refined and understood." (APA, 2017b)

Despite these types of therapies being commonly prescribed, psychologists who use CBT and NET are careful to build trust in their patient before implementing these modalities. One recommendation is to place patients in environments where their story will be received with compassion (for example, support groups for a specific type of trauma). "[Narrative] exposure is done in a controlled way," the APA says, "and planned collaboratively by the provider and patient so the patient chooses what they do. The goal is to return a sense of control, self-confidence, and predictability to the patient" (APA, 2017b). Otherwise, the patient may associate their already-traumatic experience with further distress (Van der Kolk, 1994). These works suggest that Frank's presentation of illness narrative as inherently healing is somewhat simplistic. Yet, many with trauma do express desire to communicate their experiences to others. Memoir writer Emi Nietfeld speaks strongly against these Van der Kolk's theory of trauma: "Talking about what happened again and again until it no longer tormented me sounded exactly like what I needed." (Nietfeld, 2024).

Final Thoughts:

Despite the dissent in these fields, what criticisms and counter-criticisms of illness narrative and related works do make clear to the dissertation work is that the needs of individuals dealing with illness are *individual*. Past work shows that caution should be taken when soliciting illness narratives, as the process could help or hurt any given participant. In Chapter 5 the dissertation will discuss the methods implemented to minimize this potential for harm in the IlNeS study. In regards to the work's research questions about TEO/ETRE, there are yet gaps in the field which a novel corpus may be able to explore further.

4 Annotation Work

The **Il**lness **N**arrativ**e** **S**urvey corpus (or, **IlNeS**) provides over 100 testimonials from individuals who have experienced chronic or acute illness, describing their experience. The testimonial narratives within the corpus are also new–therefore, unlike TimeBank-derived datasets IlNES had no existing gold standard annotations which could be used to derive TEO/ETRE for pairs within the text.

In this chapter, the dissertation addresses distinct challenges that emerge when producing TEO/ETRE annotations across novel data. My past works in the TDDiscourse temporal annotation effort and the NIH EpiBio time annotation project are used to explore solutions to these challenges. Finally, the approach developed for IlNeS annotation (which draws from successes in both prior annotation efforts) to efficiently produce reliable TEO/ETRE is outlined and evaluated.

Returning to the Cave:

The cave metaphorical framework proposes that text exists as an encoded projection of real-world knowledge. This projection reduces the dimensionality of that underlying knowledge and therefore, the transformation is not entirely reversible. Despite this, the aim of TEO/ETRE is to, in effect, extract a formalization of temporal properties of events within the text (*text-to-real* extraction). This might require some inference of elements not made explicit when this information was reduced and projected onto text (see Figure 29 to visualize the difficulty inherent in this process).

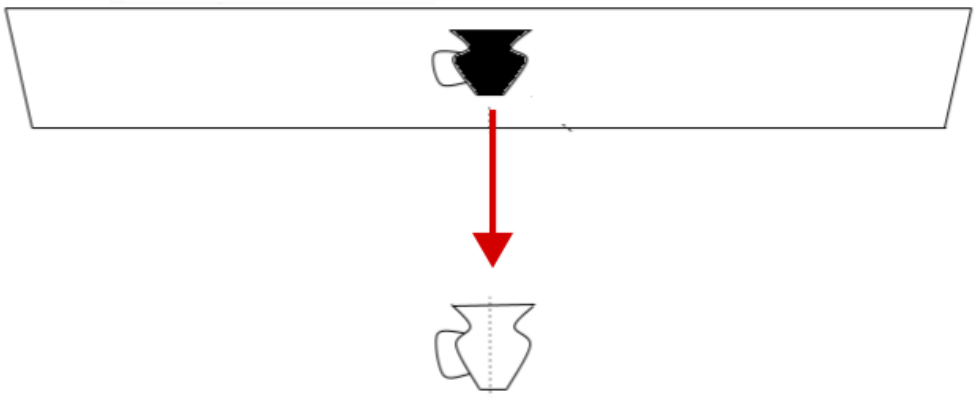

Figure 29: The limited dimensionality of the projected text makes 'text-to-real' extraction difficult without external data. Certain features, like the vase's original height and color, have been lost in translation.

Annotation is another means by which NLP researchers may engage in **text-to-real** transformation. Trained human annotators are given access to a text itself, but can rarely use more than general real-world commonsense to identify the knowledge encoded within a text. This process presents inherent challenges regardless of the annotation task; the lossy nature of real-to-text projection has a pronounced impact on TEO/ETRE annotation. Research Question 1

asks: **How can we effectively annotate new corpora for temporal elements, given the inherent challenges of this domain?**

This dissertation answers with insights from two prior works: TDDiscourse and the NIH EpiBio proprietary annotation effort. Each project had a goal to output novel TEO/ETRE corpora, and therefore the works explored the process of text-to-real projection for temporal reasoning purposes. These projects contribute methods to present the TEO/ETRE task to human annotators across distinct levels of temporal training: the TDDiscourse effort used trained annotators with experience in event and temporal annotation; NIH EpiBio annotators are trained for clinical annotation but had no direct experience in temporal annotation. Both works also identified key limitations inherent to TEO/ETRE extraction.

These results inform the experimental design of IlNeS: with it, the dissertation produces a set of *near-comprehensive* TEO/ETRE annotations, which approximate ground-truth in a way not possible in comparable corpora.

4.1 Historical Foundations of TEO/ETRE

The work is motivated by existing gaps in TEO/ETRE annotation work. Annotation within this task is traditionally done long after document creation, with trained annotators examining texts they did not write and performing text-to-real transformation. Encoded temporal information is used to predict the order in which events occurred in real time. In some cases, external information may be available–for example, the TimeBank dataset (Pustejovsky et al., 2003b) uses news articles, from which some events could be cross-referenced against other sources by later annotation teams to determine ground-truth order.

IlNeS, by contrast, represents a collection of anonymized personal medical histories. To protect the privacy of participants, external sources of personal data cannot be used to ensure ground-truth. Different approaches must be taken to ensure accuracy. This and challenges raised in prior work with other datasets motivated the IlNeS design to consider how best to produce TEO/ETRE annotations under the following conditions:

1. Annotators cannot verify the order of events in text using external sources.
2. Annotators do not have explicit training in the framework of time.
3. Comprehensive annotation of all event-pairs is required for downstream processing.

TDDiscourse explores answers to challenge condition 3) above, expanding on other annotation work done on TimeBank documents. The NIH EpiBio time annotation effort addressed the intersection of all three challenge elements.

4.1.1 Task Definition

Current TEO/ETRE uses the schema established by Cassidy et al. (2014) in TimeBankDense. TEO/ETRE's label set directly derives from Allen's interval framework of time (Allen, 1983), though the number of relations considered to be possible between two event intervals is limited to a more reasonable space. Distinct changes made to the Allen logical framework for TEO/ETRE labeling will be discussed in later sections. The TimeBankDense corpus was built from documents assembled in TimeBank. Its annotation process consists of 1) event definition, 2) time expressions, and 3) pairwise relation annotation.

Event Definition

The first step of TEO/ETRE is the extraction of temporally-significant **events**. This requires experimenters to decide what qualifies as a relevant event. In the case of TB-Dense, annotators could use the events which had already been marked by Pustejovsky et al. (2003b) as part of TimeBank annotation.

TimeBank takes texts from news sources (ex. ABC, AP Reuters, CBS, and WSJ). Some texts reflect print-published articles, and others take transcripts from live newscast segments. Annotations were completed manually by the TimeBank team, along with *time expressions*. Later annotations would expand this dataset to include TimeML date annotations for some (but not all) events (Pustejovsky et al., 2003a)[35].

For an example of a TimeBank document, see the following[36] (events in **bold**):

> Finally today, we **learned** that the space agency has finally **taken** a giant leap forward. Air Force Lieutenant Colonel Eileen Collins will be **named commander** of the Space Shuttle Columbia for a **mission** in December. Colonel Collins has been the **co-pilot** before, but this time she's the **boss**. Here's ABC's Ned Potter.
>
> Even two hundred miles up in space, there has been a glass ceiling. It wasn't until twenty years after the first astronauts were **chosen** that NASA finally **included** six women, and they were all scientists, not pilots. No woman has actually been **in** charge of a mission until now.
>
> "Just the fact that we're **doing** the job that we're doing **makes** us role **models**."
>
> **That** was Eileen Collins, after she **flew** as the first ever co-pilot. Being commander is different. It means supervising the rest of the crew in training and leading them in flight. It is, in short, the kind of management job many American women say they've had to fight for. In space, some say female pilots were **held** up until now by the **lack** of piloting opportunities for them in the military.
>
> Once Colonel Collins was **picked** as a NASA **astronaut**, she **followed** a normal **progression** within NASA. Nobody **hurried** her up. No one **held** her back.
>
> Many NASA watchers **say** female astronauts have **become part** of the agency's routine. But they still have **catching** up to do two hundred and thirty four Americans have **flown** in space, only twenty six of them women. Ned Potter, ABC News.

TimeBank annotation divides events into subtypes: *reporting*, *occurrence*, *action*, and *state*. In the example above, the following events are identified by sub-type:

1. Reportings:
   - "...we **learned** that the space agency..."
   - "...many NASA watchers **say**..."
2. Occurrences:
   - "...has finally **taken** a giant leap forward..."
   - "...for a **mission** in December..."
   - "...after the first astronauts were **chosen**..."
   - "...NASA has finally **included** six women..."
   - "...we're **doing** the job..."
   - "...**makes** us role models..."
   - "...**That** was Eileen Collins..."
   - "...after she **flew** as the first ever

[35] This process further documented by Boguraev et al. (2007).
[36] Sample text from document 'ABC19980304.1830.1636', an ABC transcript dated 03/04/1998.

co-pilot..."

- "...female pilots were **held** up..."
- "...she **followed** a normal progression..."
- "...she followed a normal **progression**..."
- "...nobody **hurried** her up..."
- "...nobody **held** her back..."
- "...female astronauts have **become**..."
- "...two hundred and thirty four Americans have **flown** in space..."

3. Actions:
   - "...will be **named** commander..."
   - "...Colonel Collins was **picked** as a NASA astronaut..."
   - "...still have **catching** up to do..."
4. States:
   - "...will be named **commander**..."
   - "...Colonel Collins has been the **copilot** before..."
   - "...this time, she's the **boss**..."
   - "...No woman has actually been **in** charge..."
   - "...makes us role **models**..."
   - "...the **lack** of piloting opportunities..."
   - "...as a NASA **astronaut**..."
   - "...**part** of the agency's routine..."

The *reporting* event sub-type is not fully applicable beyond the domain of journalism[37]. The other three are more easily generalized. According to the framework documented in Pustejovsky et al. (2003a): *occurrences* roughly cover the main predicate verb of a clause; *actions* are auxiliary verbs which support other events; and *states* represent titles, statuses, and modes of being.

> (Note: while most event mentions are nouns, verbs, and adjectives, to *be an event* is a **semantic** rather than **grammatical** status. For example, the word "that" in "That was Eileen Collins" is a cleft referent to the prior sentence, and can thus be considered to fulfill a role as a significant event in this context. The preposition "in", from "in charge", is considered to be the mention for the entire *state*.)

Though this framework captures a wide range of events reflecting a real-world timeline, most individual word tokens can be defined by this schema as event mentions. Therefore, annotators must make arbitrary decisions about the *relevance* of potential events to keep timelines within a manageable scope.

Time Expressions:

TimeBank also annotates for **timexes**, defined in Section 3.1.1. In the sample of a TimeBank document below, timexes cover the following (timexes in **bold**):

> ABC **19980304**.1830.1636 NEWS STORY
>
> Finally **today**, we learned that the space agency has finally taken a giant leap forward. Air Force Lieutenant Colonel Eileen Collins will be named commander of the Space Shuttle Columbia for a mission in **December**. Colonel Collins has been the co-pilot before, but **this time** she's the boss. Here's ABC's Ned Potter.

[37] *Reporting* events capture elements of how information about news-worthy events were gathered by reporters.

> Even two hundred miles up in space, there has been a glass ceiling. It wasn't until **twenty years** after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. No woman has actually been in charge of a mission until **now**.
>
> "Just the fact that we're doing the job that we're doing makes us role models."
>
> That was Eileen Collins, after she flew as the first ever co-pilot. Being commander is different. It means supervising the rest of the crew in training and leading them in flight. It is, in short, the kind of management job many American women say they've had to fight for. In space, some say female pilots were held up until **now** by the lack of piloting opportunities for them in the military.

Timexes are divided into subtypes: date and duration. Note that annotated timexes also include a metadata element, the document creation metadata.

1. Date:
    - "...ABC **19980304**.1830.1636 NEWS STORY..." This value is the *document metadata*, and has an annotated value of '03-04-1998'.
    - "...Finally **today**, we learned..." with value of '03-04-1998'.
    - "...for a mission in **December**..." with value of '12-??-1998'.
    - "...**this time** she's the boss..." with value of 'present_ref'.
    - "...in charge of a mission until **now**..." with value 'present_ref'.
    - "...female pilots were held up until **now**..." with value 'present_ref'.
2. Duration:
    - "...It wasn't until **twenty years** after the first astronauts were chosen..." with value of 'P20Y' (period 20 years)

Relative terms, like 'today', are linked to the document creation metadata to produce an absolute date annotation, while other terms such as 'now' and 'this time' are linked to a value called 'present_ref'. Timexes are used in some TEO/ETRE models (Mathur et al., 2021; Yao et al., 2024), but are overall limited in their contribution to final performance.

Pairwise Annotations

The last element in the TimeBankDense framework is event-pair ordering. Given two elements within the text, annotators must identify when they occurred in relative time (see Figure 30 for examples of why this process requires annotation). A major challenge in this process is scoping the boundaries of each defined event. In the TimeBank example detailed in this section, some events have clear bounds: take the event of the Space Shuttle Columbia "mission". This event mention is directly linked to an in-text **time expression** ("December"), as well as having cleanly-established **ground truth** that can be checked using external news sources. An annotator can utilize all of these indicators to situate this event in time.

This text also captures the event that female astronauts have been "made" role models. At what point can someone be said to become a role model? Is it a simple process, like selecting an astronaut as a mission commander, or does this event encompass a broad stretch of the timeline? And how would an annotator resolve uncertainties about this event's scope? This

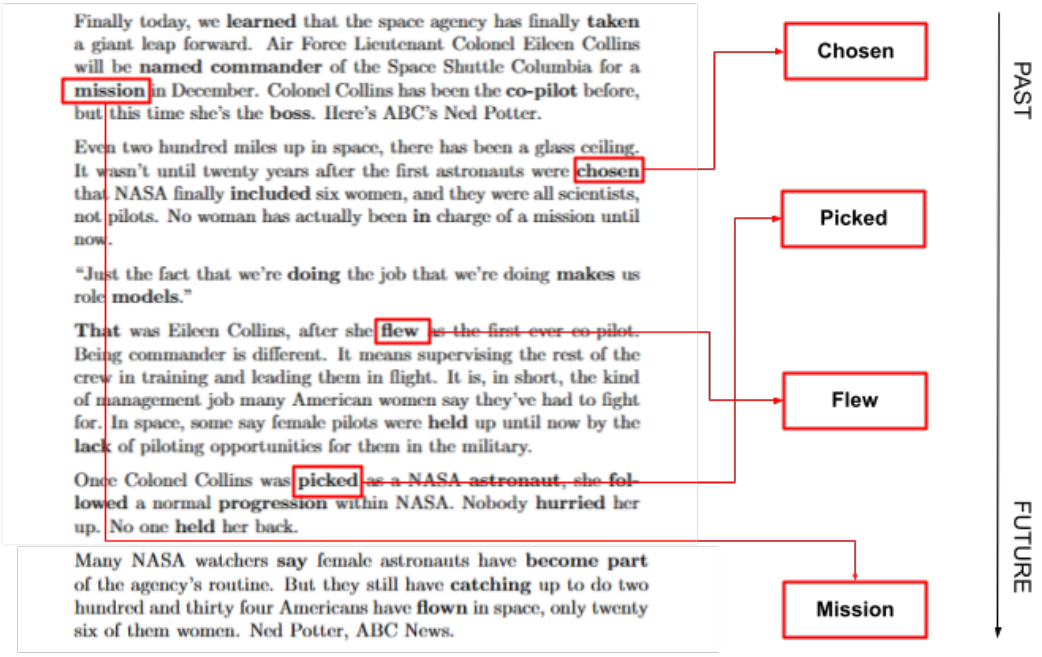

Figure 30: Events may not appear in the text in the same order as they occurred in reality.

ambiguity of event states is a non-trivial concern for TEO/ETRE, as ambiguities on the *event level* can impact prediction on the *pair level*.

In the introduction of this chapter, the dissertation noted the challenge involved with datasets where no external validation for temporal ordering exists. All TimeBank-derived corpora are advantaged compared to other datasets in this respect, as many event-pairs and temporal attributes of text have been annotated by other researchers. Further, many events described in TimeBank are of broad historical note or 'common knowledge'. However, as shown above, even breaking news documents reference events related to individual actors that do not exist in external knowledge, and which have not been addressed in other datasets.

In TimeBankDense the ambiguity of certain event boundaries is managed by annotating challenging event-pairs with a label of *vague*. This label contains no semantic data about the pair in question, but avoids the problem of annotating difficult pairs. The labels of the dataset are thus as follows:

1. **Before:** $Rel(E_A, E_B) = before$ if A starts and ends before Event B starts.
2. **After:** The inverse label to $before$.
3. **Simultaneous:** $Rel(E_A, E_B) = simultaneous$ if A and B start and end at the same point. This label is symmetric and is its own inverse.
4. **Includes:** $Rel(E_A, E_B) = includes$ if A starts before B does but continues after B starts.
5. **Is included:** The inverse label to $includes$.

6. **Vague**: TimeBankDense uses this label when human annotators feel they cannot determine the order in which Events A and B occur.

These six TimeBankDense labels derive from Allen (1983)– where the temporal value of each 'event' is a time interval–and the abridged Allen framework first formalized by UzZaman et al. (2013). This schema has inherent limitations: for example, it requires annotators to be familiar with this philosophy and perspective of temporal logic. Additionally, that TEO/ETRE annotates on the event-pair level means introduces difficulty to downstream tasks like timeline extraction. A simple method to extract timelines from texts is through *aggregation* of TEO/ETRE labels. In many cases, distinct pairwise TEO/ETRE labels may resolve to a complete, consistent timeline useful for end users. However, in cases where 1) event-pair information is **missing**, or 2) event-pair information is **internally inconsistent**, timelines cannot be resolved. See the example in Figure 31.

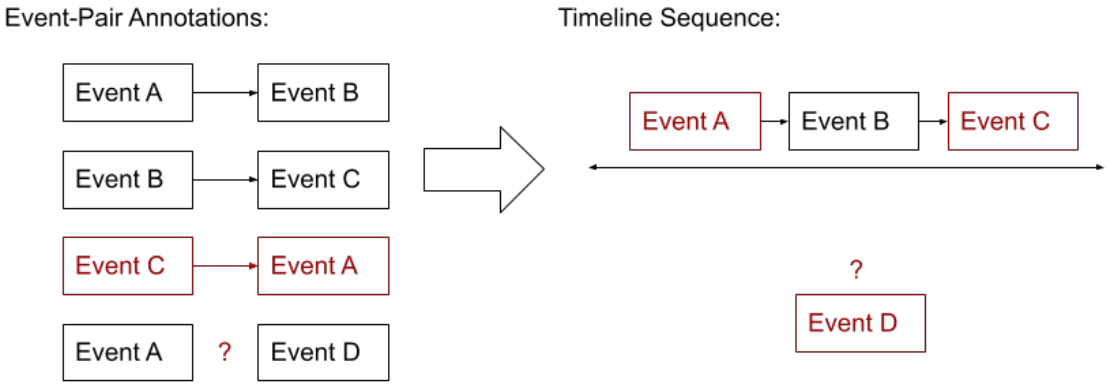

Figure 31: Timeline extraction attempt with incomplete TEO/ETRE information. Here, Events A and C are placed on the timeline but are 'wrong' due to the contradicting pair annotation. Event D lacks any annotation information, so cannot be placed.

This figure represents the case where the set of TEO/ETRE output is missing predictions $Rel(E_B, E_D)$ and $Rel(E_C, E_D)$, and where $Rel(E_A, E_D) = vague$. Due to this missing information, E_D cannot be placed on a timeline relative to the other three events. Further, there is no ordering of E_A, E_B, and E_C which fulfills all three provided predictions. It may be possible for a research team dismiss a single contradictory predicted label as incorrect. However, this is a difficult epistemological challenge with no single solution.

TimeBankDense uses the **interval approach** to train ML models for temporal reasoning. The motivation behind this choice may have been to balance the intuitive and logical aspects of time in text for ease of annotation–however, this dissertation argues that data provided by TimeBankDense dataset provides limited utility for training, for two reasons:

1. The dataset focuses on strictly short-distance event pairs.
2. TimeBankDense has a significant reliance on *vague* as a training label. 46% of its gold-standard pairs use this label (as reported in Chambers et al., 2014). Therefore, in addition to the risk of incorrect predictions, there are many E_i, E_j per document such that

$Rel(E_i, E_j)$ cannot be assigned meaningful predictions. Models trained on this corpus are likely to inherit a strong prior class bias towards this label.

These two limitations and their impacts on the utility of TEO/ETRE output are discussed more in Section 4.2. This chapter of the dissertation details two projects which attempted alternate approaches for TEO/ETRE annotation. Each proposed a different solution to improve on TimeBankDense's annotation: 1) The TDDiscourse-Manual annotation effort identified cases where the treating events in a TEO/ETRE corpus as points or as elements in a holistic sequence could successfully scaffold TimeBankDense's interval approach; 2) The NIH EpiBio time annotation work used human intuitions of time to surface more precise relationship between individual TEO/ETRE predictions and timeline sequences.[38]

In both cases, the work finds that high-reliability annotation of TEO/ETRE is more easily achieved, from both expert and non-expert annotators, with annotation schemas that use frameworks of time beyond TimeBankDense's interval approach.

4.2 TDDiscourse Annotation Schema

The project TDDiscourse by Naik et al. (2019) was motivated by a desire to produce more **comprehensive** annotations for TEO/ETRE. I was an author on this work; I worked directly with the lead making significant contributions to schema design, analyzing the needs of the corpus and where new frameworks of time could be appended to traditional approaches. I also directly assisted with annotation and adjudication. The hands-on knowledge I gained of this process informed schema design I helped develop on later projects.

A dataset is comprehensive for TEO/ETRE if the following holds true: For any pair (E_i, E_j) in a document, it is possible to infer the temporal relationship between E_i and E_j using only the TEO/ETRE annotation set. TimeBankDense does not present a fully comprehensive dataset, as it intentionally restricted its corpus to *short-distance pairs* (event-pairs whose textual mentions are within 2 sentences of one another). This limited downstream applications of TEO/ETRE–certain functions of the TEO/ETRE task flow can only be achieved through comprehensive data.

TDDiscourse sought to complement TimeBankDense by introducing the capture of *long-distance event-pairs* to the TEO/ETRE task-space. It used an iterative design approach, with four distinct implementations:

1. Base implementation: this version of the schema replicated TimeBankDense's annotation schema with few modifications.
2. Semantic ordering rules: guidelines were added for semantic relations between events which indicate reliable orderings.
3. Date-based heuristics: date-based start/endpoint-pair comparisons were added for event-pair annotation.
4. Entailment heuristics: here, the work treated event-pair annotations as portions of a larger sequence, using information about the sequence to move backward and make predictions on the event-pair level.

[38]This effort builds on EpiBio (2025b) and EpiBio (2025a), but the project itself is proprietary and not publicly available.

4.2.1 Base Implementation

The first iteration of the TDDiscourse schema directly followed the annotation method of Cassidy et al. (2014) for TimeBankDense. It first identified event-pairs from TimeBank (Pustejovsky et al., 2003b) to be featured in the final dataset, before labeling the temporal relation with human annotation. Unlike TimeBankDense, TDDiscourse annotation specifically sought to annotate **long-distance** event-pairs.

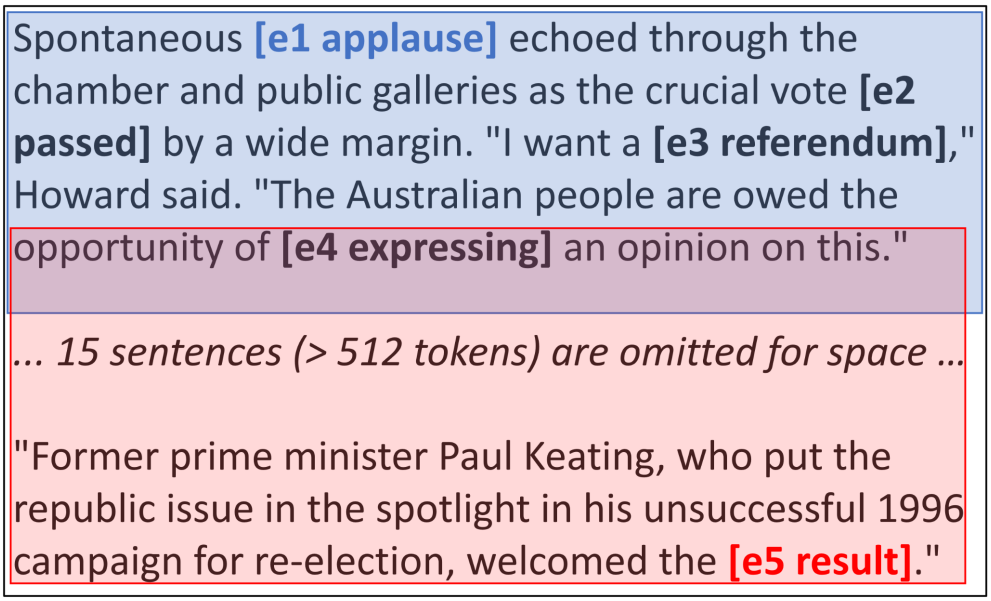

Figure 32: Different event-pairs use different cues to indicate temporal relations.

Figure 32 highlights distinct context windows present in a typical TimeBank text. Events located within the blue context window (top) form **short-distance** pairs: E_1 'applause' and E_2 'passed' are located within the same sentence, and both are 1 sentence away from E_3 'referendum'. Finally, 'referendum' is 1 sentence from E_4 'expressing'. Therefore, the following pairs are **short-distance**: (E_1, E_2),(E_1, E_3),(E_2, E_3),(E_3, E_4). Note that the pairs (E_1, E_4) and (E_2, E_4) are *not* considered to be short-distance in this schema, as their mentions are 2 sentences apart.

The choice by Cassidy et al. (2014) to restrict their annotations to short-distance pairs had obvious advantages: it allowed annotators to approach the data in a structured way and avoid the high **annotation cost** of TEO/ETRE work. Because temporal relations must be expressed as a relationship *between* two event mentions in a text, the number of possible pairs has bounded complexity of $O(n^2)$ for n events in a document.

Properties of TEO/ETRE allow all TEO/ETRE schemas (including TDDiscourse's base implementation) to reduce the number of annotations necessary for comprehensive output.

1. Because any event can be considered *simultaneous* to itself, it is not required to annotate the relationship (E_i, E_i) for any E_i in the document.
2. Because temporal relation labels are logically invertible, for any pair-of-pairs (E_i, E_j) and (E_j, E_i) for a document only require one label to be annotated.

For any document with n events, the true number of annotations required to fully build all annotated event-pair information is:

$$Pairs(n) = \frac{n^2 - n}{2} \tag{1}$$

This equation is still bounded by $O(n^2)$. The high number of required annotations is one critical challenge of annotation work. TimeBankDense, for example, has an average of 36 mentions per document (and therefore 630 total unique pairs for a text of 'average' size). Despite this, a goal of the TDDiscourse team was to approximate full comprehension for the TimeBank texts, asserting that more varied data was necessary for improved TEO/ETRE. Existing models rarely captured data outside the short-distance context range, but a study of the TimeBank corpus noted that, when linking individual events to temporal information, "for 58.72% of the event mentions, the most informative time expression is not in the same or in the previous/next sentence" (Reimers et al., 2016). This represents a substantial gap in the TimeBank corpus.

Long-Distance Pairs:

Returning to Figure 32, all pairs involving E_5 'result' are considered to be **long-distance pairs**, as E_5 is at least 15 sentences from all other events. The context window (marked in red, on the figure's lower half) covers this gap; to perform text-to-real, an annotator or ML model must surface very particular information (for example, E_5 'result' is a referent for an unknown antecedent). Structural elements of a text might help a model to make these inferences, but encoding these elements requires a fundamentally different architecture from a model trained for **short-distance** predictions.

The initial round of TDDiscourse annotation replicated the hand-annotation method done for TimeBankDense (Cassidy et al., 2014), with simple expansion of the annotation space to select long-distance pairs. This annotation round used the same labeling schema as Cassidy et al. (2014). In all cases, labels are order-sensitive; the label for event-pair (E_A, E_B) will be the logical inverse of pair (E_B, E_A). A round of adjudication allowed annotators to identify challenge cases and disagreements.

The TDDiscourse team made two modifications to the TimeBankDense annotation approach in its first implementation. First, definitions of *includes* and *isincluded* were formalized. Allen's original interval framework of time is useful for delineating boundaries between relation labels. However, it is infeasible for annotators or models to label 13 distinct relation types. Therefore, Cassidy et al. (2014) kept the underlying framework while reducing the relation label space, but documentation of the specific definitions for each label is not provided with the work.

TDDiscourse annotation work explicitly defined the *includes* label as the union of 'A overlaps B', 'A contains B', and 'B is started by A' from Allen's interval framework[39]. See Figure 33 for a visual representation of the labels TDDiscourse changes from Allen's original interval logic.

To ensure consistent labeling by adjudicators not familiar with Allen's framework, the team also used equivalent definitions based on the start- and endpoints of events:

- $Rel(E_A, E_B) = includes$ if and only if:
  - $Rel(E_{A-start}, E_{B-start}) = before$ or $Rel(E_{A-start}, E_{B-start}) = simultaneous$
  - $Rel(E_{A-end}, E_{B-start}) = after$

[39] And *isincluded* is the union of the reverse ('A is overlapped by B', 'A during B', and 'A starts B').

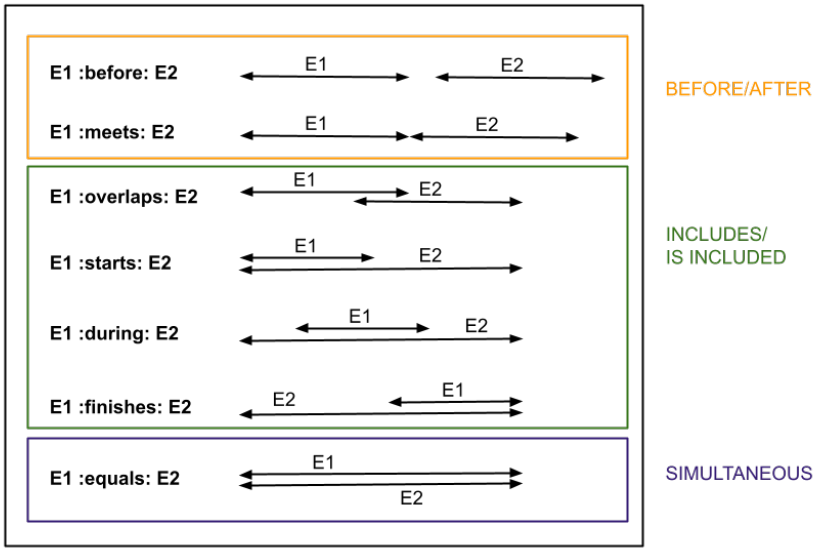

Figure 33: Conversion of Allen interval relation framework to TDDiscourse labels.

(In this definition, $Rel(E_{A-end}, E_{B-end})$ is often $after$, but this is not necessary to meet the new threshold for the *includes* label.)

- $Rel(E_A, E_B) = isincluded$ for all cases where $Rel(E_B, E_A) = includes$.

The second change TDDiscourse applied to the TimeBankDense approach was to remove *vague*, which provides no meaningful temporal data, from the label set entirely. The *vague* label was introduced for Cassidy et al. (2014) to remove ambiguities from within training data, and to ensure pairs with the remaining 5 labels could be understood to have been annotated with high confidence. However, this assessment noted that "ambiguity still remains: does VAGUE mean that no relation exists, or that multiple relations are possible?" (Chambers et al. (2014)). The over-reliance of the TimeBankDense corpus on the *vague* label for challenging pairs was significant. The corpus had produced its labels through adjudication of individual annotations from its team–of them, 83% of all event-pairs in its short-distance dataset received at least one annotation as *vague* from a human annotator. As mentioned previously, 46% of the final dataset retained the label after adjudication (Chambers et al., 2014). This provides models trained on TimeBankDense with a significant bias towards predicting the *vague* label. The TimeBankDense framework is fundamentally less comprehensive than first appears; one objective of TDDiscourse was to produce meaningful gold-standard training data for pairs which had been annotated as *vague* within TimeBankDense, along with those never annotated at all.

Results:

TDDiscourse's first iteration used a small human annotation team that produced independent annotations of the same chosen pairs. The Cohen's kappa score for Inter-Rater Reliability (IRR) reported by Cassidy et al. (2014) for TimeBankDense annotation fell between .59 and .64. This places this type of agreement in the high range of 'fair' agreement to low 'good' agreement[40].

[40] As measured using Cicchetti et al. (1981)'s standards for IRR.

This suggests that annotations in this field can have reasonable certainty, but that the tasks presents a real challenge to human annotators. (Comparatively, in TDDiscourse's first round, IRR was .48 kappa, or 'fair'.)

The lowered agreement demonstrates that long-distance event-pairs pose a greater challenge to human annotators than short-distance, with high cognitive load. Initial assessment had underestimated the challenges that TDDiscourse annotations would present: subjectivity hindered annotation and adjudication did not fully resolve this issue. This difficulty likely derives from the following:

- Many texts cover multiple topics at once. While some topics may possess an intuitive order[41], this understanding breaks down when a text discusses multiple topics in parallel.
- Distant event mentions in a text are more likely than nearer mentions to belong to separate topics.
- The relevant context for a short-distance pair is smaller than for a long-distance pair. To order two event pairs positions paragraphs apart, an annotator often has to read each paragraph between the mentions. This increases the time requirements of the task for long-distance pairs.
- Human annotators reported difficulty conceptualizing temporal order as a property of disconnected, independent event-pairs. This increased cognitive load and introduced uncertainty to the annotations.

The next iteration aimed to tackle these issues, streamlining the approach and identifying heuristics which could be applied consistently to difficult pairs.

4.2.2 Manual Coding Schema

In the second round of TDDiscourse annotation, the team developed an explicit coding schema based on existing NLP insights. This schema included:

- Because temporal similarity is necessary for **event co-reference** (discussed in Section 3.1.2), if manual annotation agreed that $Ref(E_A, E_B) = true$, then $Rel(E_A, E_B) = simultaneous$. To standardize co-reference identification, annotators used the Light ERE framework by Song et al. (2015).
- Through common-sense world knowledge, some events could be reliably observed as **causing** others. If E_A was the direct cause of E_B, then $Rel(E_A, E_B) = before$.
- Common-sense world knowledge let annotators identify events that were **prerequisite** to others. If E_A is a necessary prerequisite of E_B, then $Rel(E_A, E_B) = before$.
- When all other heuristic methods fail, E_A and E_B may be ordered based on their textual ordering. (This supports the fundamental axiom of this dissertation, that a chronological ordering of a text represents default human behavior, and it is motivated exceptions which cause a text to become disordered.)

Results:

The new annotation schema improved agreement to 69%. This is considered "good" agreement and exceeds the agreement achieved by TimeBankDense on the simpler short-distance pair set

[41] As noted in 'script theory' of social knowledge (Schank et al., 1975).

(IRR 0.56-0.64)[42]. Qualitative analysis of disagreements demonstrated a pattern also observed in Ning et al. (2018)'s MATRES work: that annotators agreed frequently (42% of total disagreements) on the basic order of events (the labels *before* and *includes* are distinct but assign the same directional order to a pair, as do *after* and *isincluded*). Disagreements were most common for the level of temporal overlap annotators believed the events displayed.

4.2.3 Date-Based Heuristics

The third iteration of TDDiscourse pulled from expanded annotations done by other researchers on the TimeBank corpus. Reimers et al. (2016) produced a date-time anchoring annotation effort for events in the original TimeBank corpus called Event-Time. This dataset has a minimum granularity of days and marks events lasting multiple days with a 'begin' and 'end' point as in Allen's interval framework. Event-Time annotations provide more comprehensive temporal information than timexes, though not every event could be given meaningful Event-Time annotation within their framework. Any 'point' (a single-day event, the start point of an interval, or the end point of an interval), could have an **exact** date-time annotation, a **partial** annotation, or **no** annotation.

For example:

- **Exact date-time annotation**: point = 1998-01-21

  Annotations use YYYY-MM-DD format. The event occurred on the date 01/21/1998.

- **Partial date-time annotation**:
  - point = before 1998-01-23

    The event occurred on some date *before* 01/23/1998.
  - point = after 1998-01-19

    The event occurred on some date *after* 01/19/1998.
  - point = after 1998-01-19 before 1998-01-23.

    The most specific type of partial annotation: this event occurred in the days between 01/19/1998 and 01/23/1998.

- **No date-time annotation**: point=unknown[43].

Event-Time provides a framework for reliable automated annotations of event-pairs across documents, both short- and long-distance. Figure 34 (below) shows the process of comparing date-time annotations for events in a pair (note: single-day events can be treated as an object whose start- and endpoint are the same date), and Figure 35 shows the full annotation pipeline.

Results:

Though not fully comprehensive (many events' dates cannot be meaningfully compared to one another), this process could be automated. This bypassed the cognitive load placed on human annotators and the potential for disagreement. This method produced annotations for both long-distance pairs and short-distance pairs which had been labeled as *vague* in TimeBankDense. This output is the dataset TDDiscourse-Auto.

[42] All Inter-Rater Reliability measured using Cohen's kappa.

[43] This annotation event is rare in the Event-Time corpus. Reimers et al. (2016) report only 0.67% of total events received no annotation.

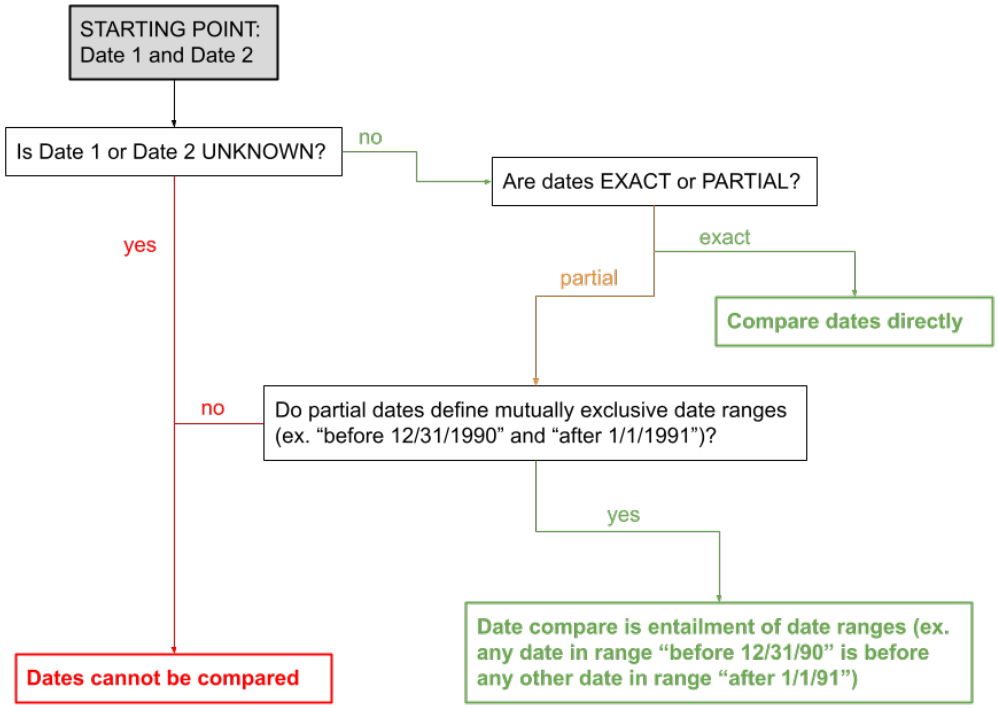

Figure 34: Flowchart for comparison of points in Event-Time corpus.

Pairs left un-labeled after this process represented the true challenge cases of TimeBank: event-pairs for which only partial absolute time information existed.

4.2.4 Entailment Heuristics

The final iteration of the TDD schema used event-pairs which had been annotated by prior implementations to derive **temporal entailments** across long-distance pairs. Cassidy et al. (2014) had noted certain event-pair relation labels that could be entailed by other relations:

- If $Rel(E_A, E_B) = simultaneous$, then $Rel(E_A, E_X) = Rel(E_B, E_X)$ for all E_X in the document.
- If $Rel(E_A, E_B) = before$ and $Rel(E_B, E_C) = before$, then $Rel(E_A, E_C) = before$.
- If $Rel(E_A, E_B) = before$ and $Rel(E_B, E_C) = includes$, then $Rel(E_A, E_C) = before$.

Note that the invertible nature of the relation labels means there are similar entailments for pairs with labels $after$ and $includes$. Entailments cannot be derived from an empty annotation set, but the entailment process is able to leverage modest manual or heuristic datasets for more efficient total annotation. Like date-time heuristics, temporal entailment can be automated.

Results:

The automated output of this implementation was evaluated using human annotators. Annotators were given 100 random pairs from the output and agreed with the automated label for

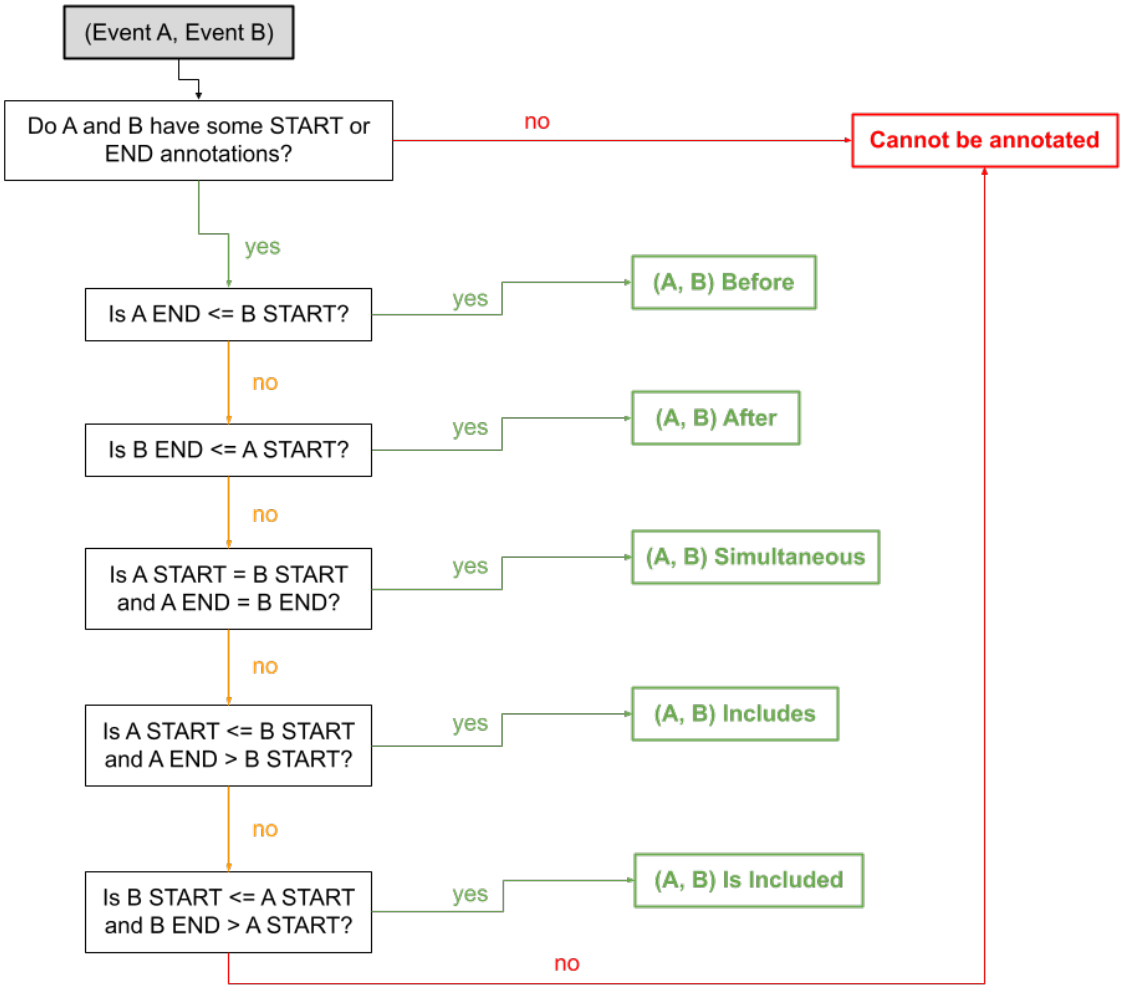

Figure 35: Flowchart for comparison of events based on begin and end points.

99% of validation pairs[44].

4.2.5 Final Corpora

The work on TDDiscourse produced two corpora: **TDDiscourse-Auto** (or TDD-Auto) and **TDDiscourse-Manual** (TDD-Man).

1. TDD-Man was obtained with human annotators applying the second iteration of the TDDiscourse schema. It consists exclusively of **long-distance pairs**. The dataset has 6,150 pairs in total, matching the size of TimeBankDense.
2. TDDiscourse-Auto was produced using the third and fourth iterations of the schema. This dataset contains 41,302 distinct event-pairs from the text across short- and long-distances. This positioned TDD-Auto as the most comprehensive TimeBank benchmark of its time.

Conclusions:

The TDDiscourse annotation effort identified the high level of challenge inherent to temporal event-pair annotation. Though the annotation team was trained and had background in temporal reasoning and event extraction, TEO/ETRE annotation imposed a high cognitive load during manual annotation. The cost of that annotation also proved a significant hurdle to the team.

TDDiscourse met this challenge by highlighting connections between event-pairs. The team found that where intuition struggled to order seemingly-disparate pairs of events, human understanding of time is *holistic*. This informs the development of IlNeS: to produce annotation schemas which are intuitive to humans, leveraging a holistic approach drastically cuts down on both annotator disagreement and cognitive load.

4.3 NIH Temporal Annotation Schema

Another major project which informed the development of IlNeS was work done with the National Institutes of Health, or NIH. The NIH Epidemiology & Biostatistics (EpiBio) research group engaged with temporal annotation for downstream integration with an in-progress patient data visualization tool. As a contractor with the EpiBio team, I organized and led the TEO/ETRE annotation effort within this larger work; I designed the annotation process and all schema as well as coding and developing tools for both automated and manual annotation.

EpiBio's long-term goals sought to produce ordered timelines for patients based on their clinical history. This required a temporal ordering model fitted to NIH's proprietary patient datasets, and new annotations to test the performance of this model. Due to the restricted nature of the data, the schema and corpus have not been disseminated outside the EpiBio department. (The schema will be discussed using only TimeBank samples or synthetic patient data within this write-up.) The work built directly from prior clinical annotation frameworks in the function domains of Mobility (EpiBio, 2025b) and Communication & Cognition, or ComCog (EpiBio, 2025a).

TEO/ETRE annotation, as established, has a high cognitive load and requires a large volume of annotations. Clinical text has distinct features compared to TimeBank (the base for TDDiscourse), and EpiBio annotators have experience scoped to *clinical* annotation but not *temporal*.

[44] As reported by Naik et al. (2019). This work did not calculate results in this instance using Cohen's kappa, and the raw data was unavailable to recalculate using kappa scores.

EpiBio's TEO/ETRE project presents a more specialized annotation framework compared to TDDiscourse; the pipeline preserved necessary features for TEO/ETRE while working within the clinical domain.

Like TDDiscourse, the EpiBio TEO/ETRE schema works to perform text-to-real translation in a way that is intuitive to human annotators. The annotation pipeline performed this transformation in steps:

1. **Event extraction**: Salient clinical events were extracted from the texts.
2. **Time Bucket implementation**: Simple event-level temporal annotations allowed the team to surface simple TEO/ETRE pairs, whose annotation could be automated similar to TDDiscourse annotation (see Section 4.2.3).
3. **Timeline visualization**: Pairwise TEO/ETRE annotations were visualized in aggregate. This approach allowed for logical temporal entailments (discussed in Section 4.2.4) to be naturally integrated into the annotation pipeline.

This approach resulted in more efficient (low time-cost) and reliable (medium annotator agreement) annotations shaped to EpiBio's proprietary clinical datasets, in a format which captured necessary TEO/ETRE information for downstream work.

4.3.1 Event Selection

The NIH EpiBio team defines, scopes, and delineates domain-specific function events within its proprietary data as part of its general focus on disability adjudication research. The team's definitions build off the World Health Organization's International Classification of Functioning, Disability, and Health (WHO, 2001), previously discussed in Section 2.3.

EpiBio has collected annotated data for four function domains: Mobility, Communication and Cognition (ComCog), Self-Care and Domestic Life (SCDL), and Interpersonal Interactions and Relationships (IPIR). Of these, Mobility (EpiBio, 2025b) was the best-annotated at when the time annotation project began.

"Mobility function events" do not match the exact format use for TEO/ETRE events. A mobility function event entity is defined as "a self-contained, well-defined description of physical functional status information [...] the definition is purposely broad to allow for normal textual variation of mobility documentation in clinical records" (EpiBio, 2025b). This requirement for self-containment reflect a specific annotation philosophy: the event entity should include all cues which indicate its annotated attributes, even when taken in isolation. This creates event entities with large spans, often covering a full clause or sentence of text.

In TEO/ETRE, models expect one-word event mentions, and cues for entity attributes are capturing using surrounding context. The EpiBio team extracted short phrases from within larger event entities using another feature of the NIH Mobility schema: all event entities contain at least one sub-span annotation known as a 'sub-entity' (EpiBio, 2025b). These sub-entities reflect elements of meaning which combine to form the full 'mobility assessment', and cover:

- **Action**: "information about activities contained within the Mobility entity [...] expressed as a verb or verb phrase, a noun or noun phrase, an inflection, a gerund, or a combination thereof."
- **Assistance**: "information about dependence on another person or object when performing an activity."

- **Quantification**: "information related to measurement values of the activity."

Sub-entities are significantly shorter than the main function entity–often only one or two words long. Quantification sub-entities are heavily specific and typically serve to modify an Actions sub-entity, but Action and Assistance have enough semantic meaning to be applicable to the task of TEO/ETRE. These sub-entities were automatically extracted from the NIH corpus to be used as events in clinical timelines.

(Note that the Mobility function dataset is tightly scoped for function "topic": though appointment records often have a similar length, some appointments are more focused on the patient's mobility than others. In the available corpus the EpiBio team identified documents with as few as 3 mobility actions, and documents with up to 40.)

4.3.2 Time Bucket Annotations

Previous annotation efforts for TEO/ETRE aimed to annotate *subsets* of event-pairs. This reduced some of the inherent challenge of the task, but the EpiBio task was motivated by a desire to produce a full, ordered timeline sequence. While entailments can expand partial information to derive new TEO/ETRE pair labels, a full timeline cannot always be produced from a partial subset of pairs. The project required comprehensive TEO/ETRE annotation to ensure full timelines.

Section 4.2 established the challenge of event-pair annotation at scale. When relations *between* events are annotated, the final selection space is:

$$Pairs(n) = (n^2 - n)/2 \approx O(n^2) \tag{2}$$

In some small documents, an annotation task with upper bound $O(n^2)$ is manageable. With documents of size $n \leq 40$, annotation is dense and overly burdensome. The EpiBio work was motivated by goals of **efficiency** and **intuition**; it required a reliable annotation schema that could be implemented quickly even on large documents. To this end, the team developed event-level **Time Bucket** annotations.

Past/Present/Future:

Time Bucket annotations are derived from event-level attributes from the NIH EpiBio Mobility Function schema, which mark entities as belonging to one of three labels: **past**, **present**, and **future**. This information is not as specific as full date-times, but there is partial temporal information encoded with this label. (The attribute in question is known within the NIH EpiBio Mobility schema as the **Timeline** attribute. To differentiate it from a textual or chronological timeline, this work will refer to the attribute as the 'NIH Timeline attribute'.)

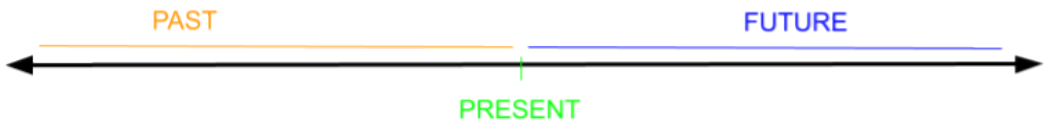

Figure 36: The range of NIH Timeline attributes, on a standard timeline.

The range of the NIH Timeline attribute can be seen in Figure 36. Examples of each label, as shown in the Mobility document write-up (EpiBio, 2025b) are listed below. Each quote represents a full mobility entity:

1. "She was independent with functional mobility prior to admit." ($Past$)
2. "She is independent with bed mobility supine to sit and sit to supine." ($Present$)
3. "Recommendation to patient: use cane for community ambulation." ($Future$)

The granularity of this attribute is insufficient for temporal reasoning (for example, two events which are both labeled as *past* could have any temporal relationship from within TEO/ETRE labels). This limits the attribute's reliability for TEO/ETRE annotation. However, this same breadth makes the attribute easy and reliable for human annotators to replicate.

Each NIH Timeline label is **mutually exclusive** along the axis of time. Based on this intuitive understanding, the following should be true for any events within a text:

- If $NTline(E_A) = past$ and $NTline(E_B) = present$, then $Rel(E_A, E_B) = before$.
- If $NTline(E_A) = past$ and $NTline(E_B) = future$, then $Rel(E_A, E_B) = before$.
- If $NTline(E_A) = present$ and $NTline(E_B) = future$, then $Rel(E_A, E_B) = before$.
- If $NTline(E_A) = present$ and $NTline(E_B) = present$, then $Rel(E_A, E_B) = simultaneous$.

This temporal logic allows some pairs annotations to be extracted automatically based on event-level attributes, achieving linear scaling with document size rather than quadratic. There was one modification made to the *present* NIH Timeline label to achieve this; in the original NIH Mobility work, short-term events which directly occurred during the appointment *and* all long-term (ongoing) states which happened to be true on that day were labeled as *present* (i.e. if a patient had been experiencing an uneven gait beginning seven months before the appointment). The difference between these two categories of event can be seen in Figure 37.

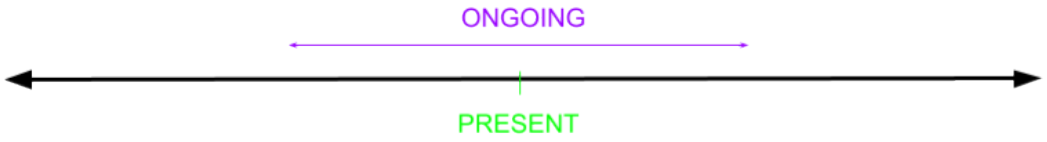

Figure 37: Two categories of events both annotated for NIH Timeline as *present*.

The EpiBio team chose to split this label for TEO/ETRE work. In prior Mobility annotation, there were two semantic categories of function event captured with this label:

1. Observations made by clinicians during the appointment, tests run, and the results of those tests. These are *short-term* events, which both begin and end on the day of the appointment.
2. Symptoms reported by the patient during the appointment. These reports nearly always reflected *long-term* symptoms of function which patients have struggled with before the appointment. It could be inferred that these events began at some point before the

appointment, and will continue to be true for some time after (even if treatment plans are successful, they will take time to diminish symptoms).

The distinction between these two categories is significant for NIH work–though the observations of clinicians are generally given more weight for tasks like disability adjudication than patient self-reports, self-reports explain essential gaps in a patient's timeline between clinical visits. The NIH Mobility annotation framework does track the source of symptom observation (i.e. whether it comes from a clinician or the patient) in another attribute: **Type**. Event entities within the category 1) above are annotated with the type label *objective*, and entities within category 2) are annotated type *subjective*[45].

Therefore, in Time Bucket extraction, a distinction is drawn between the 'true present' and 'ongoing'. The automated logic used to assign Time Bucket annotations (TB) to an event is as follows:

- If $NTline(E_X) = past \rightarrow TB(E_X) = past.$
- If $NTline(E_X) = future \rightarrow TB(E_X) = future.$
- If $NTline(E_X) = present$:
  - If $NType(E_X) = objective \rightarrow TB(E_X) = present.$
  - If $NTline(E_X) = subjective \rightarrow TB(E_X) = ongoing.$

Therefore, a rough representation of these Time Bucket categories looks like Figure 38.

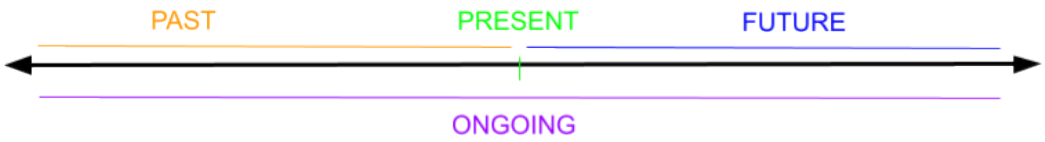

Figure 38: All four of the 'Time Bucket' categories.

Labels in this new ontology (represented in Figure 38) are not mutually exclusive–any event which is *ongoing* may start before, at the same time, or after any *past* event, for example. However, *ongoing* events will necessarily include all *present* events. Further, the introduction of this fourth label makes Time Bucket labels more intuitive to annotators, who reported finding $past/present/future$ splits to be reductive.

With this new label, the following logical entailment is added to automated event-pair extraction:

- If $NTline(E_A) = ongoing$ and $NTline(E_B) = present$, $Rel(E_A, E_B) = includes.$

Though this system cannot infer the temporal relation for all event-pairs, in practice the benefits of this shortcut are nonetheless significant for many datasets. The behavior of a corpus will depend on the type of data it contains.

[45]Very rare exceptions exist to this rule. The frequency of these exceptions was determined by the EpiBio team to be negligible in practice.

Pairs left	Documents	Percentage
0 pairs	190	39.3%
<= 100 pairs	127	26.2%
> 100 pairs	167	34.5%
Total	484	100%

Table 4: Number of pairs left per annotation document in the NIH BTRIS corpus.

In the NIH de-identified BTRIS corpus[46], the work found the a count of documents which achieved complete or significant automatic annotation using this method, listed in Table 4. Like the Event-Time annotations in TDDiscourse, this use of partial temporal information provides substantial data with little need for human involvement. Nearly 40% of NIH Mobility documents can be fully annotated through this method alone, and an additional quarter of the data sees a significant reduction of event-pairs that requires additional human annotation.

The EpiBio annotation team did not have direct experience with TEO/ETRE work. To produce fully comprehensive output, annotators had to be taught manual annotations methods for the remainder of the data. This formed the next step of the work.

4.3.3 Timeline Visualization

Key to the NIH EpiBio approach for TEO/ETRE annotation was the development of visual tools which allowed the team to interact directly with the timeline as a sequence. The event-pair annotation method successfully used in TDDiscourse annotation was considered unintuitive to EpiBio annotators; agreement was low even after adjudication, and the process was time-intensive. Traditional TEO/ETRE schema requires specialized training. However, the EpiBio visualization work, discussed in this section, demonstrates that this gap can be overcome through approaches which treat timelines as a cohesive sequence.

This approach has precedent. NarrativeTime (a project which further expands TimeBank TEO/ETRE) stated that they "replaced individual event pairs with a holistic view of the narrative represented as a timeline". In addition, researchers noted, "This solves the density problem [...] a timeline contains all the information needed for ordering all event pairs" (Rogers et al., 2024). There are two ways holistic TEO/ETRE improves efficient achievement of comprehensive TEO/ETRE: first, it allows for dense annotation of pairs through small annotator actions; second, the holistic approach reduce cognitive load by aligning with intuitive understandings of time.

The EpiBio timeline visualization tools went through multiple implementations (design mockups may be found in Appendix 10.1), but the most useful elements for annotation were as follows:

- Access to full original text alongside annotation prompts.
- Visual reference of all 5 TEO/ETRE labels.
- Visualization of partial timeline as annotations are added to sequence.

It is important for annotators that the partial timeline changes as new annotations are made. This presents its own unique challenges, as the partial timelines have to capture more distinct relationships than a full, completed timeline. Another problem of comprehensive TEO/ETRE

[46]This dataset is proprietary to NIH and can only be discussed here in aggregate

annotation, made clear in TDDiscourse work, is that difficult event-pairs interrupt an annotation workflow. It is not possible to determine which pairs will be challenging before annotation, but during the process annotators can observe the difficulty of pairs. In the EpiBio task, the team built a multi-round annotation workflow (see Figure 39) for sorting pairs.

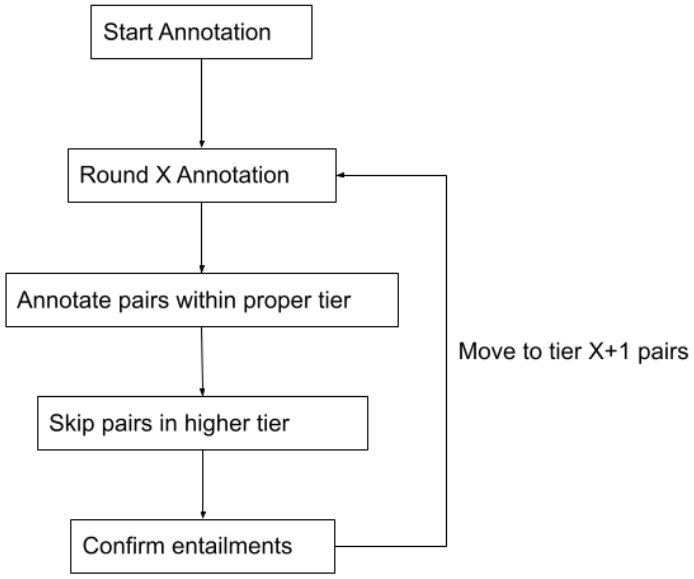

Figure 39: The workflow intended for use with Tool Version 2.

The workflow encouraged annotators to label event-pairs in rounds, starting with simple annotations and building to more difficult pairs. Difficult pairs could be skipped and set aside for later annotation rounds. At the end of each, the tool would calculate temporal entailments based on prior annotations and present them to annotators for confirmation. This reduced the overall number of annotations needed for comprehensive TEO/ETRE, and increased the chance that difficult event-pairs could be solved by entailment without direct human annotation.

The EpiBio team found that, even with encouragement, annotators did not manually skip pairs, and effectively only used one round of annotation instead of the sorted, multi-round workflow. This type of decision-making was unfamiliar even to trained annotators, and imposed cognitive load rather than reducing it. This left the team with a key insight for future corpora: annotators not trained specifically for TEO/ETRE do not benefit from multi-round annotation strategies. Finally, regarding partial timeline visualization, it became clear that the representation of in-progress timelines are inherently non-linear and therefore difficult to visualize cleanly in practice.

The type of data structure which best represents an in-progress timeline is a **Directed Acyclic Graph**, or DAG, first noted by Bramsen et al. (2006). A version of the timeline tool leveraged this as a means of lossless storage for partial timeline information on the back end, but converting the underlying DAG to a simple visual representation required some simplification. Details of timeline DAG construction can be found in Section 10.1.2 within Appendix 10.1. Overall, the following features were most useful for annotators:

- A roughly linear representation of the timeline, moving from past to future.

- Clear grouping of simultaneous event clusters.
- Separating long-span events into 'start' and 'end' points within the timeline to better represent long-term events which might overlap others.

These three features inform annotation design for the IlNeS corpus. The work with the EpiBio team (whose annotators were trained for clinical annotation but did not have experience in TEO/ETRE), directed the IlNeS project towards a simple, clear schema which could be understood even by lay survey participants.

4.4 IlNeS TEO/ETRE Annotation

This dissertation established three distinct challenge conditions that may impact production of TEO/ETRE annotations for a new corpus:

1. Annotators cannot verify the order of events in text using external sources.
2. Annotators do not have explicit training in the framework of time.
3. Comprehensive annotation of all event-pairs is required for downstream processing.

All three condition hold true for the IlNeS work. The work on TDDiscourse and the NIH EpiBio time annotation project proved that this task presents significant challenge even for trained annotators. This is especially true for corpora like IlNeS which are intended to provide comprehensive TEO/ETRE data. Therefore, the experimental design uses the survey participants themselves as TEO/ETRE annotators.

Returning to the metaphor of text as a shadow of a real object, text-to-real transformation loses some dimensionality of the real object even with refined schemas. The best way to supplement this transformation is by adding information about the real object to what can be extracted from the text. Due to the private nature of personal illness experiences, the only annotators who could ethically provide that information are the survey participants.

The IlNeS survey design will be discussed in more detail in Chapter 5. Here, the dissertation will cover how TEO/ETRE annotation was integrated into the experimental design in a way which was intuitive and effective for even wholly untrained annotators. The existing design imposed certain limitations on the annotation framework:

- Participants were unlikely to respond to surveys taking more than 1 hour.
- Participants would have no guaranteed background in TEO/ETRE annotation.
- The survey could not use existing NIH EpiBio timeline tools (for proprietary reasons).
- Surveys requiring specialized annotation tools would be less appealing to potential participants.

TEO/ETRE annotation would have to be conducted simply, quickly, and as intuitively as possible. The survey was collected through the website Qualtrics (www.qualtrics.com), a format potential participants were likely to already be familiar with, but which further limited the annotation design. The chosen method was to use questions whose answers could be submitted through freeform feedback fields. Results required some post-processing to render the outputs readable by downstream models, but the output itself would contain comprehensive TEO/ETRE data with no need for additional annotation.

Methodology:

Participants for IlNeS would be engaging with the survey asynchronously through a Qualtrics survey form. This meant that researchers could not answer questions or provide clarification about TEO/ETRE annotation after initial instruction. There are two steps to TEO/ETRE annotation on any new corpus: **event extraction** and **relation annotation**.

In traditional TEO/ETRE event extraction, words and phrases from the original text can be manually marked (ex. underlining, HTML markup). In piloting, this proved to be the most reliable method for event extraction, but basic Qualtrics surveys do not allow that functionality. The experimental design therefore opted to re-display the narrative testimony provided by the annotator alongside event extraction instructions. Participants were instructed to locate event entities from within their original testimony ("Find the [event entities] in the text related to your past personal experience with your illness, condition, or disability.") and re-write these selected events in list form within the freeform answer box.

Critical to event extraction is **event definition**. To ensure participants properly conceptualized events, instructions defined TEO/ETRE event entities for participants with the following:

> "'Actions', 'events', and 'states' [...] can include specific conditions, symptoms, or states of being (ex. 'feeling better', 'got worse'). They can include diagnoses, medical appointments, or treatments."

Additional instructions given to participants in this step were:

1. "Consider the narrative from your own perspective only."

   This instruction cues participants to select relevant, important events from their own testimonials. In piloting, over-defining the concept of 'event relevance' confused participants. This phrasing tested well for influencing participants to select relevant subsets of the total events.

2. "Do not list events you did not write about in [the original text]."

   It is important for TEO/ETRE that every event can be linked to a *textual mention*.

3. "Do not list 'hypothetical' events (i.e. things you wish had happened, things you plan to do, things you were told could happen but did not)."

   In NIH EpiBio work, annotators provided feedback that hypothetical events (common in clinical text as clinicians discuss prognoses, goals, and possible side-effects of treatment) lack a real temporal value and therefore cannot be easily annotated.

4. "Try to sum up the event in one or two words, ideally using the same language as from [the original text]. For example, if you believe an important event in the text is, 'I was diagnosed last year", you might write 'diagnosed'."

   TEO/ETRE expects short event phrases, and must be linked to a *mention* from the original text.

The event relation annotation for IlNeS builds directly from NIH EpiBio insights: **it is more intuitive for human annotators to conceptualize of a timeline as a single cohesive sequence than as an aggregation of event-pair labels.**

The limits imposed on the IlNeS annotations due to the one-hour Qualtrics format mean that the survey does not have time to elicit a full set of pair-level annotations from untrained annotators. However, pair annotations can be extracted trivially from any ordered timeline sequence. By asking participants to arrange events in a timeline sequence, the work leverages intuitive

understandings of time to overcome the lack of participant training. Further, it extracts comprehensive annotations without the usual $O(n^2)$ annotation time.

In the Qualtrics survey design, the list of event entities marked by participants is re-displayed, and the following instruction given:

> "Using the actions and events from [event entity list], arrange the event using bullet points in the order that they actually happened in real time. This may, or may not, match the order that they appear within the narrative. If an event does not appear to correspond to any real time (i.e. it did not actually happen), don't put it on the timeline."

To ensure the full complexity of a TEO/ETRE timeline could be preserved by participant annotation, the survey design gave instructions for three timeline features: **events before/after one another**, **events taking place simultaneously**, and **events which include other events**. The instructions show participants different examples of "how two events may be arranged in bullet points":

> If: Event A takes place before Event B
>
> - Event A
> - Event B
>
> If: Event A takes place after Event B
>
> - Event B
> - Event A
>
> If: Events A and B take place at the same time
>
> - Event A + Event B
>
> Suppose Event A takes a long time (months, years). Identify the point this event starts and ends when constructing your timeline. Ex:
>
> - Event A Start
> - Event B End

Like the event extraction step, participants are given a freeform answer box on Qualtrics in which to write this timeline sequence.

Results:

Of the 106 participants in the survey, 12 did not provide annotations for events or timelines. This suggests the instructions were not entirely clear to those participants. Many of the 12 used the freeform response fields to add continuations to their original illness narrative. The dataset preserves this input, but did not generate new TEO/ETRE annotations during post-processing to avoid introducing ordering errors. In the remaining 94 texts, post-processing was able to produce standardized timelines.

The most frequent error type in the IlNeS TEO/ETRE annotation came during the event extraction step. Most participants did not use exactly the wording and phrases of their original sample text; this is a specific requirement of TEO/ETRE modeling which participants were completely unfamiliar with. Some participants wrote in new events not mentioned in their original text, but which clearly formed an unspoken background of their written narrative.

Others used wording that was similar but not identical to mentions in the original text. Further, events as listed often had to be reduced to grammatical head words of the phrase to better match TEO/ETRE input expectations.

Post-processing handled these errors in distinct ways:

- Events which were not identical to the original text but had distinct lexical features (ex. event written as "feeling lost" but the closest phrase in text is "felt lost") are converted to their closest textual equivalent.
- Events which had distinct semantic links to a specific phrase in text (ex. event written as "overwhelmed with pain" but 'overwhelmed' does not appear in text). If a positive semantic connection can be made to a phrase in the text (in the case above, "the pain feels heavier than anything" was deemed the closest equivalent), that is used as the event mention.
- Events which were too different from mentions in the text (ex. the event "feel alone" for a text which had no direct emotional statements. While the loneliness of their experience was clear subtext across the work, no wording from the text was semantically similar). In cases where the link between a listed event and potential mentions were too tenuous, the event was dropped from the final timeline.

IlNeS post-processing also standardized cases where event text was non-unique (ex. the word "effect" is common in the clinical domain and may appear multiple times in a text). If no indication had been made by the participant of which mention the event corresponded to, the first instance in the text was chosen to avoid biasing the data towards the research hypothesis[47]. Multiple listed events with the same text are linked to distinct mentions in the text, again by order of appearance.

Errors in the event-pair relation step could not be discerned as easily, as IlNeS cannot compare the participants' 'gold-standard' annotations against ground-truth. The IlNeS corpus assumes all orderings provided by a participant are correct. The only noticeable errors in this step were formatting errors. Though the instructions provided standard markers for long-term events (START and END markers for the begin/end points), simultaneous event clusters (+), and events following others in sequence (placed on different bullet points in the list), participants often used their own nonstandard markers. When the nonstandard markers were clear about annotation intent, post-processing converted these to standard format.

One interesting error type was participants marking an event with the "start" marker, but not providing a corresponding "end"[48]. These were long-term conditions which were likely to be true until the present moment (similar to *ongoing* events from the NIH EpiBio work), so post-processing added corresponding "end" events for all such events in a simultaneous cluster representing the present moment.

The methodology of the IlNeS TEO/ETRE ordering allowed a majority of participants to annotate their own texts for temporal ordering. Table 5 counts the number of participants who include a certain TEO/ETRE label type in their annotations.

[47] Research Question 3 asks if there certain types of narrative which exhibit temporal deviations disproportionate to current class biases of ordering models. It would be possibly to theoretically choose optimal instances per text which better fit our research hypothesis that narrative types do behave differently.

[48] It is possible the phrasing of the original instructions made this element of long-term events less clear to participants in a way not caught during piloting.

Total Documents	Any Labels	Before/After	Simultaneous	Includes/Is Included
106	94	88	53	31

Table 5: The frequency of documents with certain annotation labels (distinct from the frequency of annotation labels across documents).

These counts serve as lower bounds for the overall "comprehensibility" of a TEO/ETRE annotation type. It is always possible that a type of annotation was understood by a participant without their narrative happening to use that TEO/ETRE label. The low rate of timelines using split START/END events, for example, could be due to confusion over events overlapping (consistent with NIH EpiBio annotator feedback) or a low rate of overlapping events overall (consistent with TDDiscourse label distributions).

3 participants directly relayed confusion about the directions given. One complaint was primarily about technical limitations of the Qualtrics form ("I was think [sic] how much better it would be to have a large box where I could see what I had already written without scrolling.") and two regarding the schema ("I apologize if I did the previous section of this study incorrectly with the bullet points. I was unsure exactly how to arrange everything."; "These instructions are absolute convoluted rubbish. The average participate [sic] will have no idea whatyou [sic] want."). Overall, these represented a small proportion of the 106 survey participants.

Broadly, survey results suggest that the majority of participants had at least a basic understanding of TEO/ETRE after instruction, despite the known complexity inherent to the task. Their real-world knowledge of the specific events in their testimonials leaves the IlNeS annotated timelines as close approximates of ground-truth for future experimentation. The annotated timelines allowed for comprehensive event-pair extraction in post-processing (the contents of which will be discussed further in Chapter 5).

4.4.1 Contributions

TEO/ETRE annotation is an inherently difficult task, even for trained human annotators. Schemas for works being annotated after document creation require significant intervention to reach 'good' Inter-Annotator Agreement, and the process is costly (as shown with TDDiscourse). Advanced tools can help visualize time in a more intuitive way for annotators, but come with their own development and design concerns (see NIH EpiBio work).

The IlNeS corpus demonstrates that 1) creators of a text have a specialized expert knowledge of the text's real properties which can substitute for and potentially exceed trained TEO/ETRE annotation; 2) it is possible to communicate the needs of TEO/ETRE data to untrained annotators within one hour in an asynchronous text format; and 3) timeline sequence annotation greatly exceeds the efficiency of event-pair TEO/ETRE, and is highly intuitive to a lay audience. Though there remained some points of conceptual confusion that required post-processing, the IlNeS work was highly successful in producing complex temporal data for the new corpus, and presents a model by which such annotations can be collected for new corpora efficiently in the future.

5 Illness Narrative Survey

This dissertation asks the following research questions:

- **RQ1. How does the purpose of a narrative change the order in which temporal events are presented?**
- **RQ2. What can we observe about the link between temporal positioning in text and intended effect from short-form standardized narrative text?**
- **RQ3. Are there certain types of narrative which exhibit temporal deviations disproportionate to prior class biases of ordering models?**

Narrative analysis suggests that the intention with which an author writes a text (called simply 'authorial intent') will change the order in which events are presented in-text. This ordering may deviate enough from texts which are commonly used to train temporal ordering models that adjustments may be required.

To test this, the dataset must be annotated directly for both temporal ordering and intention. Illness narrative is an ideal domain for this work; it is tightly scoped (therefore reducing confounding variables like topic, register, and genre constraints) but retains a large body of theoretical study on the internal motivations of text types. This chapter presents the corpus **IlNeS** (**Il**lness **N**arrativ**e** **S**urvey). IlNeS is annotated for textual and chronological timelines, and is partitioned by the intention label **NarraType**. The chapter outlines the process of data collection for this corpus, the data itself, and perform analysis of the corpus to answer the research question. It also explores the following secondary questions:

1. What can the testimonials in IlNeS tell us about existing theories in the field of illness narrative? Does our work support or refute existing assertions?
2. What do these testimonials tell us about how patients communicate with their clinicians?

I am the principal investigator for this survey work; I designed the survey protocol, recruited participants, ran the survey, processed the data, and performed qualitative analysis.

5.1 Motivations

The end goal of the dissertation is to incorporate directed, human-centered analysis of text into algorithmic machine learning. This dissertation begins by building a corpus from that human-centered perspective. Illness narrative is a field with a large body of qualitative analysis performed by knowledgeable experts, but which has not yet been joined to a mathematical approach. IlNeS contributes a dataset designed for downstream ML integration, bridging these two perspectives (as will be discussed further in Chapter 8).

The dissertation constructs a new dataset for this work, as none previously existed to fill this exact space. Existing TEO/ETRE datasets often draw from a single domain with fairly uniform 'authorial intent'. Pustejovsky et al. (2003b)'s TimeBank sources events from journal articles about breaking news–the primary intention of news is to **inform** their imagined audience of current events. It could be argued that in some cases, journalistic bias introduces additional intention types to a text. Sub-genres of journalism (like advocacy journalism or opinion journalism) may explicitly consider their primary goal to be **persuasion**. But texts used by TimeBank are sourced from mainstream news sites considered to be generally reputable[49] and

[49] ABC, the Associated Press, CNN, Reuters, Wall Street Journal, and more.

not situated within a sub-genre with persuasive intent. Accusations of reporting bias within these spaces is highly subjective, and annotations of that type are not included in the original corpus. This tight scope and lack of labeling for texts which might diverge from standard journalistic intention make TimeBank and derivative corpora unsuited to test the dissertation hypothesis. Specialty datasets for TEO/ETRE exist in other domains, including medicine (ex. L. Zhou et al., 2006). But such corpora share the underlying problem of TimeBank: there is no objective method to separate texts in the scoped domain to reflect distinct intention types.

The field of **illness narrative** suggests a possible solution to these problems. Prior works have sorted narrative corpora into distinct goals or intentions (see Frank, 2013, Bury, 2001). However, as discussed by Strawson (2004), the majority of work in illness narrative is qualitative rather than quantitative. The scope of existing data is impractical for downstream applications like ML modeling. New data was necessary to rigorously evaluate the dissertation research questions. Annotations of illness narrative provide another benefit: though this topic is not strictly tied to the main research question of narrative intention, the dissertation is interested how the typical patient approaches interactions with a clinical provider. Patient-clinician communication remains a critical concern of modern-day healthcare, and the attitudes a patient brings into that conversation influence how effectively a clinician can provide care. The intention with which a patient approaches clinical interactions could serve as a measure of the overall level of trust they are invested in the interaction (for example, if a patient's speech patterns in a clinical interaction behave most like persuasive narratives, it suggests friction exists in their healthcare experience). The IlNeS corpus work simulates such interactions to surface insights which may be of use to the field of health communications.

5.1.1 Experimental Design

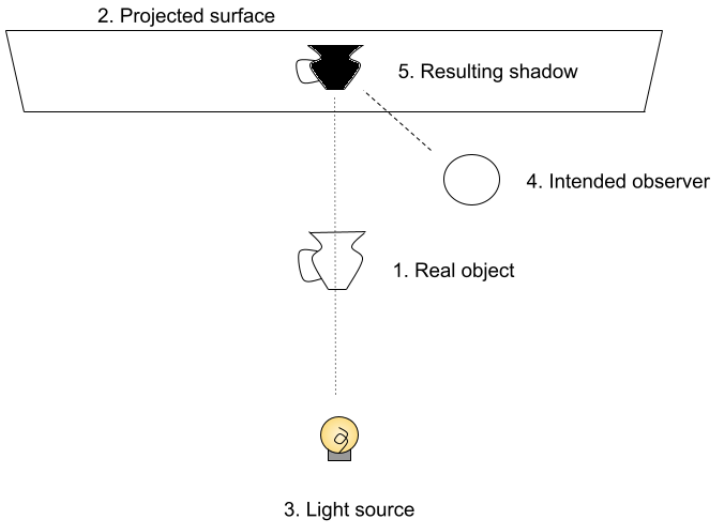

Figure 40: The framework of text and context.

Returning to the framework of text and context introduced in Chapter 2 and visualized in Figure 40, there are external factors related to any text which the author does not control. To apply this metaphor to illness narrative testimonials, participants of this survey cannot change

the events that happened to them (1. the Real Object), the format and time limits of the survey (2. the Projected Surface, which captures a distorted projection of the real object), or the their testimonial's audience (4. the Intended Observer). Each factor influences what the observer sees in the final text (5. the Resulting Shadow). In the metaphor of casting a shadow, the only element survey participants can directly control is the framing and structure of the text itself (3. the Light Source).

The IlNeS corpus is designed to capture the interaction between elements 3) and 4) of the text-context framework–elements of the text which change due to an author's expectations of their audience. See Figure 41 to understand how the initial audience variable influences intention/style, and how that influences time and temporality within the text. The survey strictly controls the first step in this chain, the Intended Observer of the text. Therefore, the resulting text timelines can be understood as a measurable consequence of steps along this chain.

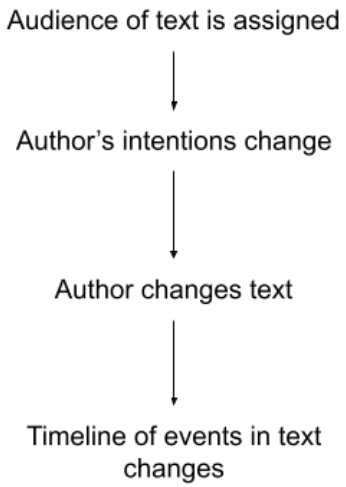

Figure 41: How the proxy audience may influence the order of events in text.

The corpus captures the following:

- An independent proxy variable for each participant's simulated audience. The work associated this variable with a marker of authorial intent: "narrative type" or **NarraType**.
- Full testimonials about the participants' experience with illness, disability, or other medical conditions.
- Comprehensive TEO/ETRE annotations per testimonial. The process by which the survey obtains these annotations from participants is described in Section 4.4.

5.1.2 NarraType Typology

The typology of the survey's NarraType variable builds on existing work within the illness narrative field to more rigorously scope the attribute of authorial intent or impact. Authorial

intent is most often studied in what are considered "high-skill" works of narrative–lay writers may not recognize the concept when asked to identify it within their own work. Therefore, IlNeS used proxy prompts to prime participants for a particular type of narrative. The scope of NarraType labels were defined through synthesis of existing illness narrative typologies.

Chapter 3 discussed different typologies for authorial intent which exist within the field of NLP. A common problem among these type frameworks is overspecialization to a given task. IlNeS supports its taxonomy with detailed citations of existing and accepted work in the field of **illness narrative**. The work combines three typologies within the field: Frank (2013); Hydén (1997); Bury (2001) (all discussed in Chapter 2). The aim of the new taxonomy was to contribute a general NarraType framework that was applicable for multiple domains. The taxonomy therefore seeks to be less granular than previous works and to clearly partition its rhetorical categories.

The IlNeS NarraType labels are detailed below. For each, this dissertation will show an example from within the IlNeS corpus. Note that these chosen examples are intended to illustrate *archetypal expected* behavior for the NarraType; the IlNeS corpus was designed to measure the varied behavior of participants when asked to produce distinct narrative types, and individual testimonials from within it may therefore differ from the exemplar.

1. **Factual Narratives** present a factual retelling of events. This may include the medical facts of an illness, its progression, and its treatment, and the purpose is didactic. This sub-type is built using criteria from the concept of the *restitution narrative* (Frank, 2013) and the *contingent narrative* (Bury, 2001). Factual illness narratives are most likely to arise from a need to educate others about a condition, though they may be produced for other reasons.[50]

   > *Ex.* (from Participant 2276) "I would like to articulate a little of what I have been experiencing, since to the outside it may not make much sense. My condition is something that is not well-known to most people, and it does not always appear so apparent, yet it has an immensely significant effect on my everyday life. How unpredictable it is is one of the key problems. On some days I am virtually who I used to be, and other days I find it difficult to do even simple tasks such as getting out of bed, concentrating or following conversations. I cannot just power through it with willpower, it is my body imposing constraints I have no control over. That uncertainty renders planning problematic. I may say that I will go out or assist with anything, but I must cancel at the last minute due to my symptoms. People always suspect that I am not reliable yet it is my condition that defines what I can or cannot do. It even influences my self perception. The thing is that I did not give energy or stamina a second thought before. I need to be careful now, and even there I usually feel exhausted or frustrated because I cannot keep up with other people. That has been a hard emotional pill to swallow, since I do not want to feel like a burden. Meanwhile, I have grown to be different in a manner other than negative because of this condition. I have learned to take my time, to celebrate the little successes and to be more understanding of others who have some struggles that no one sees. It has not been simple, but it has helped me to be strong and patient which I

[50] Bury theorized that patients benefited in an emotional manner from factual re-tellings of narrative; by recounting events, patients could make sense of their own illness experiences (Bury, 2001). The IlNeS work makes the assumption that this act of catharsis is a byproduct of the testimony, and not the deliberate intention behind it.

believe I would not have otherwise learned. And so you may not see my illness but you know it is always with me- it will influence the way I move in the world. And I do like it when people listen and attempt to understand because such support is really a difference maker."

2. **Persuasive Narratives** make some case to an audience using personal experience as example. Often the specific aim is to re-frame societal beliefs about illness or to argue for material change. This sub-type builds off of–but is more broad than–the *moral narrative*, which identifies a specific case where an author feels the need to present themself as a moral actor who did not cause or deserve their own illness (Bury, 2001). It also draws from the *narrative as strategic device* (Hydén, 1997).

   Ex. (from Participant 0848) "Thank you so much for considering allocating funds/resources into researching better treatment options for celiac disease. Currently, the only treatment for celiac disease is a gluten-free diet. While gluten-free options have become more commonplace, it is still very difficult to live a normal life and avoid ever being glutened. An additional treatment option, in cases of accidental glutening, would go a long way for others with celiac disease. Aside from avoiding gluten, having an autoimmune disease comes with other issues as well. I regularly experience nausea, rashes, fatigue, and body aches. I do my best to perform my daily tasks at work and at home, but it's a constant struggle. I think more research into autoimmune diseases in general would help a lot of people. According to my research, one in 100 people have celiac disease, so 1% of the population. Beyond that, between 5-8% of the population have some type of autoimmune disease, and this number may even be higher considering how many people have likely not been diagnosed. When this is put into perspective, it becomes a significant portion of the population. Please consider allocating additional funds/resources into researching better treatments for celiac disease and other autoimmune diseases. It would help a significant portion of the population live healthier lives and be more productive workers that are less of a burden on the healthcare system."

3. **Emotional Narratives** re-frame the experience of illness as a story where the author acts as an agent within the narrative. Often there are lessons learned and some positive takeaway from the illness itself. This builds from the theoretical concept of a *quest narrative*, which proposes that the act of telling a story about illness can directly empower an individual dealing with illness by creating a sense of agency despite external circumstance (Frank, 2013). The *core narrative*, which posts self-affirmation as the ultimate goal of such texts (Bury, 2001). Unlike other illness narrative sub-types, existing theoretical work supports the idea that the main beneficiary of this narrative type is the author themself, despite the fact that such texts are often constructed for audiences.

   Ex. (from Participant 1340) "Although it may seem like something minor, something that will not limit you greatly, this disability will affect your life in ways you don't understand right now. You will need to steel yourself, to do your best, and to accept your limitations in order to be happy and at peace. I have been dealing with this disability since I was a teenager. At the time I began to experience it, I didn't really understand it. I don't think that the medical staff I dealt with did a very good job explaining how a 'limited range of motion' would change me, which is why it is extremely important that you ask as many questions as you need to, as many times as you need to, in order to make truly

> informed decisions about your life. Because if you don't ask, the doctors may not tell you. They may not think to tell you. It is unfortunately your job to make sure that you have enough information. Another frustrating thing about this disability is that it sets me right in the middle between 'not at all disabled' and 'legally physically disabled'. If I wanted to, I could pursue things like a handicapped parking sticker, or some kind of disability payments. I am limited by my pain and my physical body in the kind of work that I can do. At times, I need a cane to walk. At times, I am in great pain. Stairs are difficult for me. But, I do not require a wheelchair. I can still complete my ADLs (activities of daily living) without help and with only minor modifications. So, while I feel crippled, I also feel (mentally) that it could be a lot worse. Finally, you will receive attention and comments from friends and strangers. People have no issue asking me 'what is wrong with my leg' or assuming that I am injured. It feels terrible and makes me feel like a spectacle. I hate it. I wish it is something that I could rise above and feel more comfortable about, but I cannot. That is the most difficult thing, to me, about being disabled. I do not see why my body is something others feel so comfortable speculating about or commenting on. Try to approach this with good humor, or understanding, or pity, or whatever you need to in order to be more okay and comfortable with it. Remember, it could be worse."

The NarraType typology preserves elements of most subtypes from major existing illness narrative systems. One significant subtype absent is the *chaos narrative*, which Frank defines as a heavily non-chronological narrative type in which motives of sense-making and self-reflection break down completely (Frank, 2013). This type of narrative is theorized to be driven by external factors–see the discussion of factors which may impact memory in Section 3.3.2–rather than rhetorical motivation. It is therefore beyond the scope of the primary dissertation hypothesis and excluded from NarraType. The three categories of purpose defined by this typology ("to explain", "to persuade", "to express emotion") are *generalizable* and can easily be applied to texts outside illness narrative, though work in the dissertation will focus on this domain.

The choice to keep focus on illness reflects its versatility as a topic of narrative. Labov proposed that certain subsets of human experience prompted raw and honest expressions of intention; illness, or "death and the danger of death"[51] was one which he reported elicited less-filtered testimonies from participants compared to other topics (Labov et al., 1997; Labov, 2013). This expectation motivates the design of IlNeS as a study of illness testimonials. The dissertation acknowledges limitations of the NarraType typology to analyze this domain–there is an argument that deeper, richer analysis could be achieved by 'parsing' a text into sections with distinct NarraType attributes. Similar to how discourse and rhetorical parsers mark discrete 'acts' throughout a text, it may be that intention information is best captured on the sentence or segment level. This dissertation lays groundwork for future explorations of text using that framework, but begins its analysis on the document level. It assumes that testimonials can be generated in full from one primary intention, and that this motivation will hold true on some level throughout the whole text.

Major confounding factors are discussed in Section 5.5. Other potential compounds include poor memory influencing participants' retelling of events, cultural elements of language, and biases towards certain types of narrative. These confounds are unavoidable given the recruitment

[51] Interestingly, Labov associates this topic with the emotional state of 'thrill', while Frank and peers tend to describe a more contemplative state from participants.

process, though IlNeS mitigates the potential impact by soliciting testimonials from as broad a subset of real individuals dealing with illness as possible. (Recruitment details are discussed further in Section 5.2.1.) However, conclusions from this study are scoped to those participants, and may not generalize for demographics beyond those recruited for the study. Note also that participants are not informed they will be annotating their testimonials for temporal order until after they have constructed the full testimonial[52], to avoid participants shaping their narrative for this later task.

5.2 Survey Setup

5.2.1 Setup and Recruitment

Recruitment for this survey was done in two phases. The first iteration of the study used physical flyers and online advertisement to organize one-on-one video calls, with the intention of primarily running the survey as an interactive interview. The second phase connected to participants through the research platform Prolific (app.prolific.com), with a link to an asynchronous survey hosted by Qualtrics (qualtrics.com). The two-phase process was implemented due to slow recruitment in physical recruitment. Few reached out for the physical survey, and fewer scheduled interviews; the vast majority of testimonies in the IlNeS corpus were provided during the second survey phase. Therefore, this work will primarily discuss the implementation of the second survey phase and the format of instructions as given to this group, with occasional notes about Phase 1 implementation.

The use of Prolific allowed the research team to be connected to interested parties who met the inclusion criteria (adult English speakers, located in the United States, past experience with chronic illness or disability). Prolific recruitment contacts a set number of random potential participants until all survey 'spots' are filled. This ensures the participant group should otherwise reflect a random distribution of the population. Participants were given a link to an external Qualtrics survey, and (after completion) a verification code to access payment. The research team made a deliberate choice not to link survey results to participants' unique Prolific IDs[53]. This prevented researchers from re-identifying participants after survey submission or accessing their demographic information. This was done to minimize the chance of identifiable participant information being linked to survey answers.

Participants (in both phases) were given an one hour to sign the consent form and complete the survey. Phase 1 participants used the full hour, but most in Phase 2 finished early (with a Phase 2 median completion time of 33 minutes).

The consent form was emailed to participants of the video-call survey and included in the asynchronous version for download, with a prompt to upload their signed copy for receipt. In the video call format, the researcher could walk through the form to ensure that participants had read and understood the details. This was not possible with the asynchronous survey, nor was it possible to guarantee participants had signed and returned their consent form before proceeding to the full survey. However, this form (approved by IRB) contained thorough details about survey risks, emergency resources, response storage, and future use of data, as well as contact information for the Office of Research Integrity and Compliance. Instructions were clear that video-call and asynchronous participants should read the document thoroughly and keep a copy for their records. In cases where participants did not provide a signed copy of their

[52]Except to the extent required to obtain fully-informed consent.
[53]Prolific offers this feature but it is not required in order for participants to receive payment.

consent form but still answered the survey questions, the team opted to pay the participant for their time but omitted the submitted survey response from the corpus records.

Prolific collection occurred in rounds of 20-25 new participants at a time, to allow flexibility in the final corpus size based on ongoing analysis. In a survey consisting of a single session, 'drop-outs' are not as significant a problem as with long-term studies. However, there were cases where participants left the survey without fully completing it: 1) participants followed the link to the Qualtrics survey but opted not to sign the consent form, exited before seeing survey questions, and 'returned' the survey to Prolific; 2) participants uploaded valid consent forms but exited and returned the survey before answering questions; 3) participants did not upload a valid consent form, but followed the survey to completion. These cases slowed data collection, but had a larger impact on the NarraType distributions of usable surveys.

The design of IlNeS sought random, even distribution of NarraType prompts per partition. This could be enforced in Qualtrics by setting the survey to show 1 of 4 random prompts for each new user accessing the survey. The expected distribution would be roughly even per NarraType. However, 'drop-outs' sometimes skewed collection; if multiple potential participants happened to be given the same prompt before dropping out[54], that NarraType would be underrepresented in resulting data. Collecting data in rounds allowed us to modify the random prompt assignment per round, based on which NarraTypes were required to even out corpus partitions. Because the research team did not directly recruit participants, all Prolific participants were anonymous even to researchers, and Prolific had no insight into the NarraType selection, the team is confident that the matching of NarraTypes to participants remained as random as could be ensured.

Note: Participants were assigned identifiers at random from numbers in range 1-10000. All survey data was stored only under these anonymous identifiers, and references to specific testimonials through this dissertation will use the identifier. The random selection was done to obscure the order in which responses were collected and further reduce the chance of re-linking responses to consent forms.

5.2.2 Collection

Questions were presented to participants in Phase 2 through remote access to an asynchronous Qualtrics survey. There was a trade-off to this approach, as (unlike in the live interviews) participants who did not understand the survey instructions could not reach out for clarification. This did help ensure that the presence of researchers did not unintentionally influence the results of the survey[55].

To test the hypothesis that there exist observable correlations between temporal ordering and intent, the IlNeS work uses **NarraType** as an independent variable and examines the patterns of **temporal deviations** that emerge for each.

Therefore, the design of the survey required the following: 1) *prompts* designed for specific intents, 2) annotation of *events* within the narrative, and 3) a *chronological arrangement* of those events. To better validate this survey design, the protocol adds 4) a *freeform feedback* section for rich qualitative data.

[54]Note that most drop-outs occurred before participants could have seen their assigned prompt, so it is likely random chance that some prompts were over-represented among drop-outs.

[55]In the video-call surveys we could conduct, the research team clarified for participants in cases where they were confused by instructions on how to *format* their answers, but avoided providing details on the *content* those answers should contain.

The majority of data collection was done through a Qualtrics-hosted survey. Qualtrics collected e-consent forms by providing a download link to the full consent form (discussed in more detail in Appendix 10.2.1) and a space to upload the new document after providing an e-signature, as shown in Figure 42.

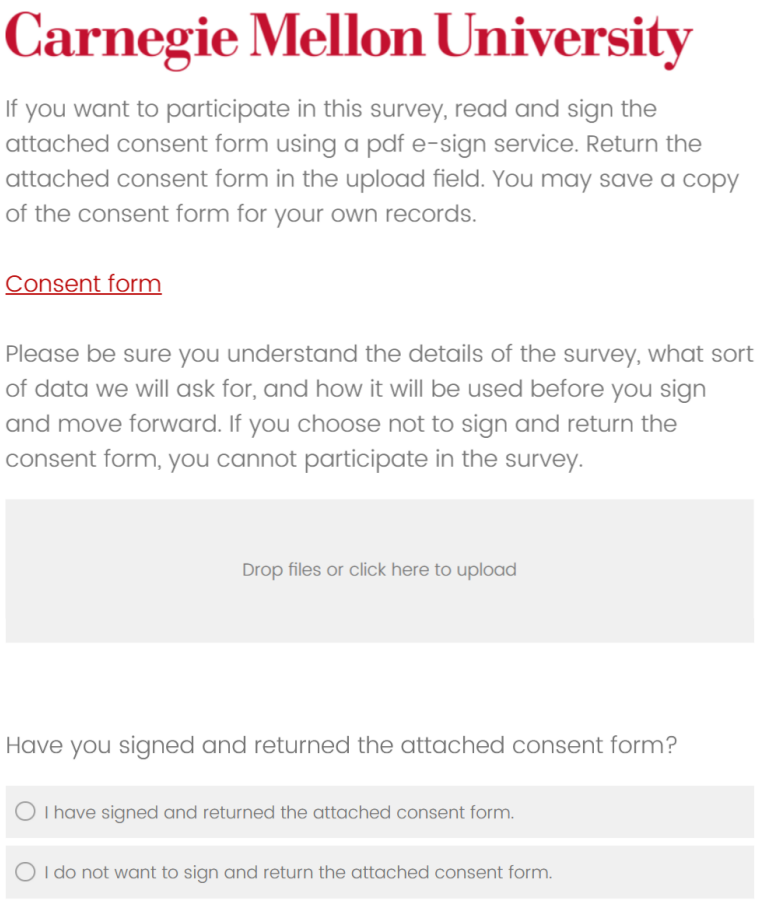

Figure 42: The Qualtrics form for collecting e-consent.

Participants who selected the option "I do not want to sign and return the attached consent form" were not allowed to participate further in the survey. All participants who selected "I have signed and returned the attached consent form" were allowed to proceed. It was not possible to validate that participants who picked this option did sign their e-consent form; the research team had to validate later that the uploaded a document and that this document had some signature confirming consent.

Instructions for Testimonials:

The first challenge of the IlNeS work was to overcome the artificiality that survey protocols often induce in participants (discussed by both Labov (2013) and Sneller et al. (2023) in their studies of oral narratives). Labov considered formal interviews of the type used by IlNeS "more guarded, and less interesting" than "[the] style of speech that is superior to all others–from the

linguistic point of view–which we call the vernacular" (Labov, 2013)[56].

IlNeS does seek to observe the natural behavior of participants. However, the work does not consider formal writing training to be at odds with instinctive behaviors of intention. The work is also already placed within a domain considered to provide honest and natural testimony. Therefore, the formal survey format was considered unlikely to cause significant confounds for results.

The IlNeS survey provided participants with the following initial instructions, verbatim, for their individual testimonials:

> "Write your personal account of how your illness, disability, or medical condition has impacted your life. Please try to imagine that you are sharing this story in a real-life scenario. You will be given more details on the specific audience to imagine next page.
>
> If you are dealing with more than one distinct medical condition, focus the following questions on one condition you feel you can give a better accounting of your experiences with.
>
> As this is a written account, you can make edits to what you have written within the time limit, though you do not need to. You do not need to use the full time limit (twenty minutes) if you believe your work is complete before then.
>
> In this question, you will be given a scenario to imagine with a goal to achieve, but you may write your personal account in the way you believe will best meet the goal of the prompt. There is no 'correct' answer to this question–we are looking to understand how people like yourself respond to these types of scenarios.
>
> IMPORTANT: Names, locations (ex. hospitals), specific dates (ex. date of a surgery), and some diagnostic or treatment details are considered 'personally identifiable' information that could be used to de-anonymize your data. Do your best to avoid using these details in your survey response, though our team will also remove any personally-identifiable details left in the survey data before dissemination to protect your privacy.
>
> In addition, we do not require that you include any specific details of your illness experience in this survey. You may choose to omit any personal details for any reason, without penalty. In this survey, we encourage you to include details only if 1) you are personally comfortable with our research team having, retaining, and sharing these details with other researchers in an anonymized format, and 2) you feel it is notable for the specific scenario given."

The priorities of the survey design reflected in these instructions were as follows:

- Participant privacy:

  A key concern in the construction of the survey was to respect the sensitive nature of personal illness. An individual's medical history, if made public, could cause them material and emotional harm in their day-to-day life. Therefore, the research team sought to minimize the chance of collecting unique details which could re-identify an anonymous participant. The team also emphasized participants' agency over what they shared in

[56]This is a long-standing attitude among linguistics–Vygotsky (1986) considered 'natural' registers to behave most closely to thought.

testimony, encouraging participants to omit all details (identifiable or otherwise) they were not comfortable disclosing anonymously to a broader audience.

- Natural reflection of human experience:

  As previously discussed, the survey format may elicit unnatural patterns of text from human participants. In the case of IlNeS, the research team had concerned that participants may attempt to tailor responses based on assumptions of what the study was 'looking' to find in data. Assurances that participants may best help the study by providing honest and unfiltered responses can mitigate some degree of self-censorship.

- Capture elements of written text:

  The written text format differs from speech in that it is asynchronous (i.e. the reader does not perceive the text until after the writer has constructed it). This allows for some elements of format–like revision and editing–which are not possible in spoken language (speech self-repair[57] corrects, but does not replace, mistakes in a contiguous 'text'). This work is scoped to written text, and therefore captures elements within this domain such as editing.

- Standardization of format:

  The annotation and analysis of clinical events becomes more difficult when multiple medical conditions were being treated simultaneously. It is not uncommon for individuals with chronic conditions to have other, co-morbid conditions; to produce a standardized testimonial format, participants are encouraged to focus on a single condition.

Audience as Proxy:

Because the goal of IlNeS was to use NarraType as an independent variable and analyze its impact on testimonials, the work required a survey format that collected authentic answers while still directing participants towards a specific narrative intention. "Intent" is a difficult quality to classify in natural human text. The work differentiates "intention" from "deliberation"; even subconscious moves to shape language for the benefit of the reader qualify as falling within *authorial intent*. However, this subconscious element of intent makes it difficult to elicit directly. Lay writers may not feel confident in reproducing techniques of different intention types if asked to do so directly.

Therefore, the work sought to solicit this authentic production of intent through *audience design*. Bell (1984) presents audience design as a primary factor in responsive stylistic choices within text: "Style is essentially speakers' response to their audience. In audience design, speakers accommodate primarily to their addressee." This phenomenon has also been observed by Hall-Law et al. (2022)'s project on video diaries post-COVID, where participants gravitated to distinct but consistent styles across all entries based on their expectations of audience (which, due to the video diary format, were simulated by participants). "Many participants in the Lothian Diary Project orient to the task of diary making with an elevated sense of seriousness and formality," the work noted, while others "evoke the style of social media vloggers" (Hall-Law et al., 2022). IlNeS builds on this principle, extrapolating that participant engagement with even imaginary audiences will influence the text and style of the testimonial.

IlNeS elicited this differences of intention through the use of **imagined audience scenarios**. Participants were instructed to imagine their narrative was being told to one of four audiences. These proxy prompts correlated for the underlying element of intent; Section 6.1 demonstrates

[57] See Levelt, 1983.

that these proxy prompts successfully oriented participants towards approaching tasks with factual, persuasive, or emotional intent.

An example of how a participant might be given a proxy prompt is shown in Figure 43. The prompt which appears on this page is chosen at random before participants begin their survey. Each participant is shown only one prompt. A freeform text box allows participants to write their testimonial within the allotted 20 minutes, as shown in the sample below.

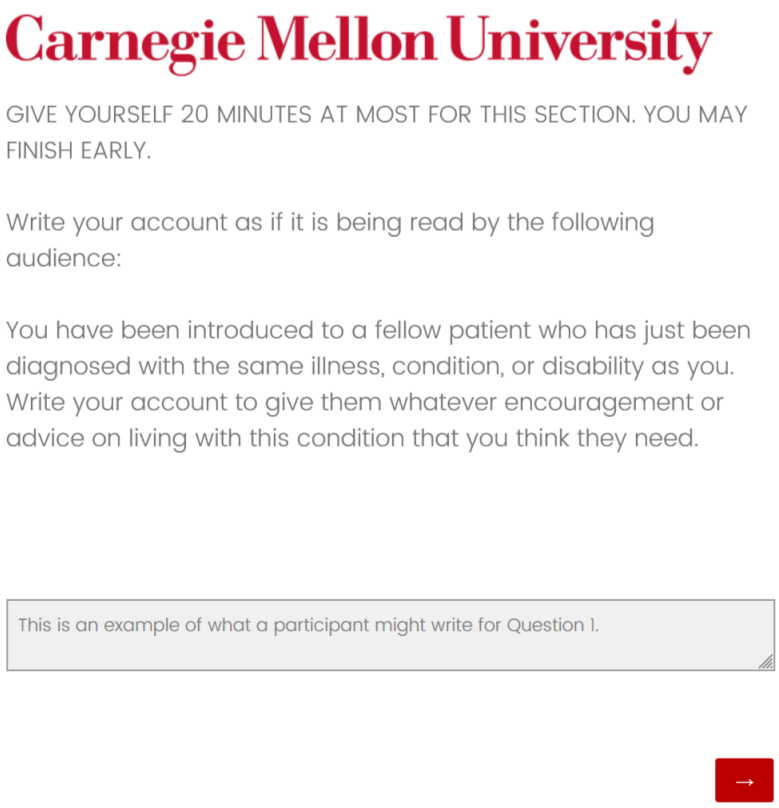

Figure 43: The Qualtrics page presenting participants with Question 1 and their audience proxy prompt. In this case, the prompt shown is for the Emotional NarraType, and an example of an answer has been provided in the freeform text box.

There are four proxy prompts, one corresponding to each NarraType, which may be shown to the participant at this stage. Prompts are as follows:

1. "A friend or acquaintance has never heard of your illness, condition, or disability and does not know anything about it. Write your account as if explaining to them your experiences with this condition."

   This scenario proxies for the **Factual** NarraType.

2. "A policymaker (for example, a politician, lobbyist, or researcher) is considering allocating funds/resources into researching better treatment and accommodation options for your illness, condition, or disability. Write your account as if convincing them to do so."

This scenario proxies for the **Persuasive** NarraType.

3. "You have been introduced to a fellow patient who has just been diagnosed with the same illness, condition, or disability as you. Write your account to give them whatever encouragement or advice for living with this condition that you think they need."

   This scenario proxies for the **Emotional** NarraType.

4. "You have switched general practitioners and are talking to your new GP for the first time about your medical history."

   This scenario is not a proxy for a true NarraType. Rather, it is used to explore how patient discussions with clinicians align with the other three intention modalities. This is called the **Clinician** NarraType label through the rest of the work.

   An example of a Clinician-type testimonial from IlNeS (sample chosen to be illustrative of the expected archetype of Clinician testimonials, other texts may vary):

   > (From Participant 8372) "I appreciate you taking the time to go over my medical history with me. Since this is my first visit with you, I'd like to give you a clear picture of how my condition has impacted my daily life. One of the biggest challenges I face is managing fatigue and physical limitations due to my condition. I work as a [POSITION] in a hospital, where I'm responsible for [DETAIL], [DETAIL], and [DETAIL]. Some days, I can push through without too much difficulty, but on other days, I feel like my energy is depleted much faster than it should be. This affects my ability to be as physically active as I once was, and I've had to adapt by delegating tasks more strategically and pacing myself throughout the day. Pain and discomfort are also recurriing [sic] issues. Whether it's a dull ache that lingers or sudden flare ups, it impacts my ability to focus and sometmes [sic] makes daily responsibilities more exhausting. I do my best to manage it with lifestyle adjustments, but I'd like to explore any recommendations you might have for long term relief or management strategies. Beyond the physical aspects, my condition has also influenced my mental well being. I priortize [sic] my mental health, but there are moments when frustration or stress builds up, especially when I feel like my body isn't cooperating with what I need to do. I've devloped [sic] coping mechanisms, like breaking tasks into smaller steps, utilizing AI tools for efficiency, and ensuring I get proper rest, but I want to be proactive in managing this aspect of my health as well. My goal is to continue leading a fulfilling and productive life while managing my condition in the best way possible. I'd appreciate any insights you have on treatment options, pain management strategies, or lifestyle changes that could help improve my day to day experience."

Every participant was given one of the four above prompts, neatly partitioning the IlNeS data by NarraType. Participants were not shown proxy prompts for NarraTypes they had not been assigned; though there was potential in examining if asking a participant for two testimonials about the same illness experience would surface changes in text based on NarraType, this method seemed unlikely to match natural intention patterns (as participants might be influenced by their first testimonial while writing the second).

Event Selection:

As discussed in Section 4.1.1, a significant challenge in building datasets for the task of TEO/ETRE is **event selection**. In TEO/ETRE models which predict relations on an event-

pair level, this problem may bias a model towards different kinds of pairs; in IlNeS annotation work, participants produce comprehensive (or nearly-so) sets of pair annotations by constructing timelines as holistic sequences.

It is impossible to capture the full granularity of a real-world timeline, as nearly all semantic objects in a text can be 'events'. All TEO/ETRE annotation efforts make some judgments on which events are considered significant. In IlNeS, participants were used as the arbiters of event significance. The exact details of IlNeS's implementation of event selection and TEO/ETRE annotation are found in Section 4.4. As a reminder of the exact instructions given:

> "Go through the answer you gave in Question 1 (displayed below). Find the 'actions', 'events', and 'states' in the text related to your past personal experience with your illness, condition, or disability. For this question, consider the narrative from your own perspective only.
>
> Actions, events, and states can include specific conditions, symptoms, or states of being (ex. 'feeling better', 'got worse'). They can include diagnoses, medical appointments, or treatments. Do not list events you did not write about in Question 1. Do not list "hypothetical" events (i.e. things you wish had happened, things you plan to do, things you were told could happen but did not). Try to sum up the event in one or two words, ideally using the same language as from Question 1. For example, if you believe an important event in the text is, 'I was diagnosed last year', you might write 'diagnosed'."

Here, Figure 44 covers the input area of the page shown to participants for Question 2 (extraction of events from the text). The input provided for Question 1 (the "testimony") is shown to participants within this page.

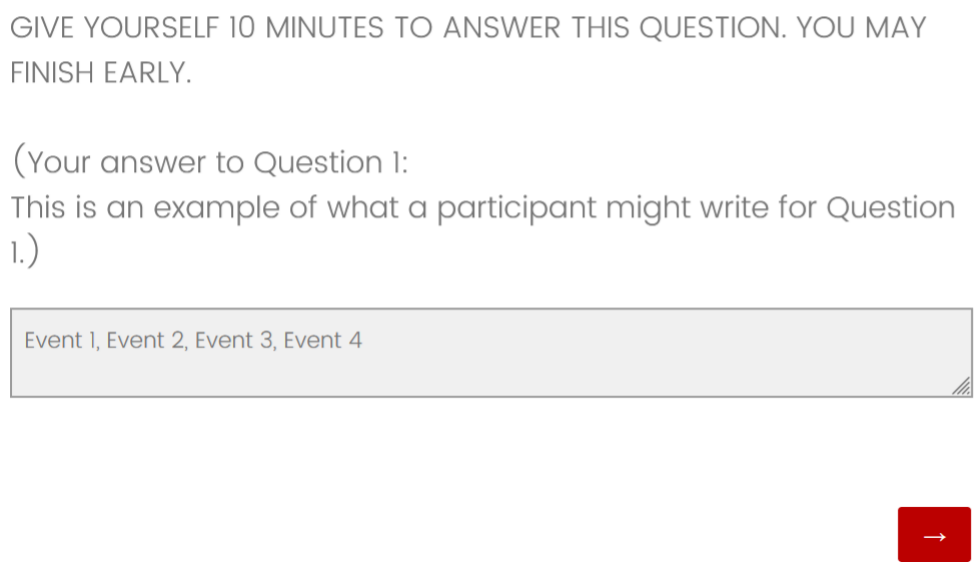

Figure 44: The input field for Question 2, event extraction. The text input for Question 1 has been displayed alongside the new input field.

Note that nothing in the Qualtrics format enforces specific selection of event mentions from

within their original testimonial; at times, researchers needed to perform post-processing of data to match listed 'events' with mentions from the text.

Timeline Annotation:

After the extraction of a list of events from within the testimonial, the IlNeS work requires participants to build a timeline orienting the event list in chronological time, as discussed in Section 4.4. Justifications for the design of this element of the survey may be found in that section; this section of dissertation lays out the final form of the survey as given to participants.

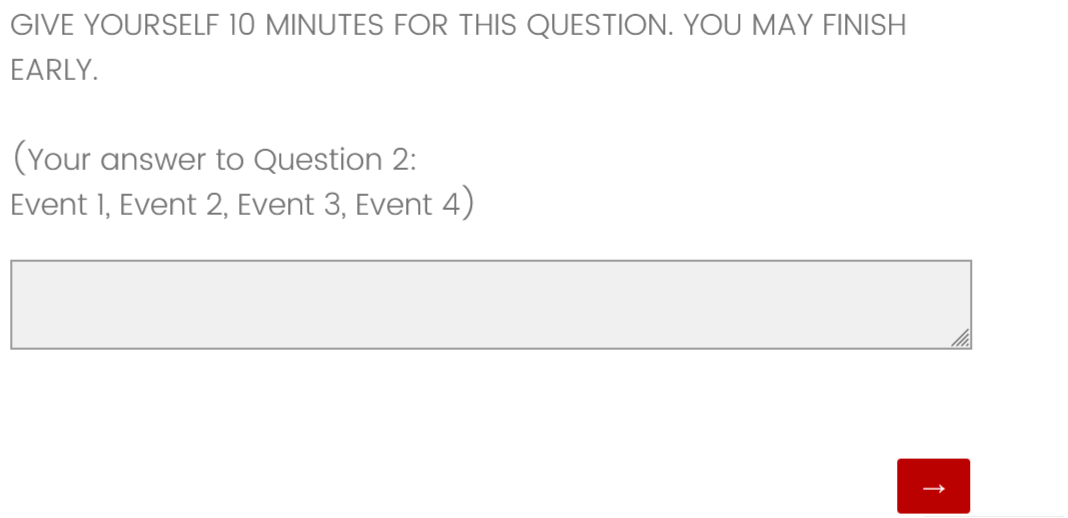

Figure 45: The input field for Question 3, timeline annotation. The text input for Question 2 (event extraction) has been displayed alongside the new input field.

Question 3 was framed as follows:

> "Using the actions and events from Question 2, arrange the event using bullet points in the order that they actually happened in real time. This may, or may not, match the order that they appear within the narrative. If an event does not appear to correspond to any real time (i.e. it did not actually happen), don't put it on the timeline.
>
> To show how two events may be arranged in bullet points.
>
> If: Event A takes place before Event B
>
> - Event A
>
> - Event B
>
> If: Event A takes place after Event B
>
> - Event B
>
> - Event A
>
> If: Events A and B take place at the same time

- Event A + Event B

Suppose Event A takes a long time (months, years). Identify the point this event starts and ends when constructing your timeline. Ex:

- Event A Start

- Event B End"

Figure 45 shows the input field provided to participants for Question 3 (timeline annotation). Again, the Qualtrics survey format does not allow enforcement of the annotation format, and researchers often had to perform post-processing to produce standardized timelines.

Freeform Feedback:

To better understand the role of intent in text generation and support observations about illness narrative as a genre of text, IlNeS collected free-form responses from participants about their thought process in the survey. This question was open-ended to avoid biasing participants towards any specific disclosures. Trends within this data therefore reflect noteworthy commonalities between participants.

The wording used to collect qualitative participant observations was as follows:

> Describe, in as much or as little detail as you can, your own thought process as you wrote your answer to Question 1 [the testimonial prompt]. This section does not have to be a complete or proof-read short essay–any text which describes how you were thinking during the writing in Question 1 is useful to our study.
>
> Take however much time you want, and you can move to the end of the study if you have nothing you'd like to share with us.

5.3 Methodology of Analysis

After collection, the IlNeS dataset covers 106 autobiographical illness testimonials. These testimonials are linked to full timeline representations of relevant events in true chronological order, from which all event-pair annotations can be automatically extracted. In this section, the dissertation analyzes the resulting timeline data to answer Research Questions 1-3.

5.3.1 Timeline Taxonomy

To explore RQ1 (**How does the purpose of a narrative change the order in which temporal events are presented?**), it is necessary to analyze the timelines produced by IlNeS annotations and compare to the ordering of these timelines within the corresponding text. This dissertation defines a *temporal deviation* as a phenomenon of non-chronological ordering within a text (more broad than just two events in a pair which are ordered non-chronologically); in this section, different types of temporal deviation are identified and quantified across distinct NarraTypes are quantified, which allows the work to track trends and draw broader conclusions about NarraType behavior.

This project intends to quantify the impact of NarraType on timelines; this requires a taxonomy which allows the work to differentiate timelines based on their chronological properties. The framework used by this dissertation begins with an intuitive understanding of time. Take as example the following example Timeline 1, in Figure 46, from the IlNeS corpus. This sample is a **chronological** timeline. All chronological timelines should be considered equivalent for the purposes of ranking temporal deviations.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"diagnosed"	"metformin"	"diet and exercise"	"endrocrinology"	"well controlled"
Textual:	"diagnosed"	"metformin"	"diet and exercise"	"endrocrinology"	"well controlled"

Figure 46: Timeline Type 1: Chronological. Example from Participant 0251.

Timeline 2 in Figure 47 is a **mostly chronological** timeline. Small, local-level deviations have been introduced to the original temporal sequence. This deviation type may be caused by grammatical constructions common to to English. Consider the common English construction "X happened because of Y"; this statement flips the temporal ordering of X and Y, but is not necessarily indicative of specific narrative intent. Next, timelines with many temporal deviations are examined. The taxonomy identifies multiple ways these deviations can appear, and seeks to determine how different types of temporal deviation ought to be ranked.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"struggle"	"diagnosis early"	"rehabilitation, medication, and therapy"	"study"	"get a job"
Textual:	"diagnosis early"	"struggle"	"rehabilitation, medication, and therapy"	"study"	"get a job"

Figure 47: Timeline Type 2: Mostly chronological, with the deviation highlighted. Example from Participant 1616.

	Position 1	Position 2	Position 3	Position 4
Chronological:	"screening"	"early detection"	"left untreated"	"diagnosed"
Textual:	"diagnosed"	"left untreated"	"early detection"	"screening"

Figure 48: Timeline Type 3: Inverted. Note the ordering of events in each sequence. Example from Participant 5382.

Timeline 3 (Figure 48) is fully **inverted**. By contrast, Timeline 4 (Figure 49) is more randomly chaotic and **disordered**. Both can be intuitively understood as types of temporal distortion, but it is not immediately clear which should be considered "more" disordered. One intuition would propose that the inverted timeline is further from chronological, as every individual pair in an inverted timeline is non-chronological. However, there is an argument that an inverted timeline still preserves order; if inversion is applied to the timeline again, it becomes chronological. Meanwhile, disordered timelines require many operations to restore a chronological sequence.

This taxonomy is useful from a qualitative perspective. To quantify the relative deviation of each timeline, the work requires metrics for which the following are true:

- All chronological timelines have the same metric score.
- No timelines outside the chronological timeline type have this metric score.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"grieve"	"lonely"	"support" (1)	"manage"	"doing more"
Textual:	"manage"	"enjoy the little moments"	"doing more"	"grieve"	"living a full life"
	Position 6	**Position 7**	**Position 8**		
Chronological:	"support" (2)	"enjoy the little moments"	"living a full life"		
Textual:	"support" (1)	"lonely"	"support" (2)		

Figure 49: Timeline Type 4: Disordered. Note that some event mentions use the same text but represent distinct mentions and events. Example from Participant 2768.

- Different types of high-deviation timelines should be captured in distinct ways across metrics.

5.3.2 Rank Correlation Coefficients

To extract standardized metrics of timeline deviation, the work conceptualizes the pair of timelines T_{chr} and T_{txt} as sequences containing the same elements in different orders. Pairs within (T_{chr}, T_{txt}) that have high levels of similarity should be treated as chronological, while pairs whose sequences differ reflect temporal deviance. This framework is similar to the task of **rank correlation**, where sequences are ordered (or "ranked") for a certain criteria, and two systems are meant to provide similar ranking outputs. **Rank correlation coefficients** are one measure for this underlying similarity. This section explores three such metrics for rank correlation:

1. Spearman's footrule.[58]
2. Kendall's distance.
3. Cayley distance.

These metrics are considered standard for web sorting (Dwork et al., 2001) and other forms of sequence evaluation (Fligner et al., 1986; Irurozki et al., 2016). To utilize ranking metrics, the work first represents T_{chr} and T_{txt} as numerical vectors (call these T'_{chr} and T'_{txt}), where each event in the timeline is numbered according to its chronological "rank". Therefore, for a timeline T_{chr} of size n, T'_{chr} will be the vector $[1, 2, ...n]$ and T'_{txt} will be a vector containing the same ranking numbers in an order according to textual appearance.

(Note that one distinct property of chronological timelines is that two events can be temporally ranked as a "tie" if they occur simultaneously. Therefore, in the case that two or more events are tied for temporal position i, each event in T'_{chr} (and their counterparts in T'_{txt}) is labeled as having value i.)

Spearman's footrule:

Spearman's footrule is a non-statistical representation of rank distance, with lower scores correlating with high similarity. It is traditionally represented by the equation;

[58] Another metric used for rank comparison is Spearman's rank correlation. These metrics use similar formulae and are likely to surface redundant information about the corpus. This dissertation uses the footrule and omits rank correlation.

$$d_S(T_{chr}, T_{txt}) = \sum_{i=1}^{n} \|T'_{chr}(i) - T'_{txt}(i)\| \tag{3}$$

Because the timelines in IlNeS do not have a set length (the number of events per timeline is decided by the individual participant), footrule scores can't be compared directly without normalizing. Traditional Spearman's footrule normalization divides by $n^2/2$ for sequence of length n (Dwork et al., 2001).

Kendall tau distance:

Kendall tau distance (or Kendall distance) $d_K(T_{chr}, T_{txt})$ counts the minimum number of pairwise adjacent transpositions are required to change sequence T_{chr} to T_{txt}. This can be calculated algorithmically, and is mathematically equivalent to the *total number of pairs for which the order differs between sequences*:

$$d_K(T_{chr}, T_{txt}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \bar{d_K}(T_{chr}(i), T_{chr}(j)) \tag{4}$$

Here, $\bar{d_K}$ is a characteristic function with the following values:

$$\bar{d_K} = 1 \iff Pos_{txt}(T_{chr}(i)) > Pos_{txt}(T_{chr}(j)) \tag{5}$$

$$\bar{d_K} = 0 \iff Pos_{txt}(T_{chr}(i)) < Pos_{txt}(T_{chr}(j)) \tag{6}$$

$Pos_{txt}(x)$ returns the position of the provided item in the sequence T_{txt}. In simple terms, the characteristic function which is summed to produce Kendall's distance counts the number of cases where ordered pairs across the original sequence are disordered in the new one. (This value is similar but not identical to a minimum **edit distance**.) For TEO/ETRE, it is equivalent to the full count of pairs in a document with labels $after$ or $isincluded$. This means that the distribution of labels within the document (and therefore the ideal distribution of predictions for a TEO/ETRE model) will change with the Kendall's distance.

To normalize Kendall's tau, the raw score is divided by its absolute maximum potential value[59], which is $(n * (n-1))/2$ for a document of size n.

Cayley distance:

Cayley distance $d_C(T_{chr}, T_{txt})$ counts the minimum transpositions total needed to convert the two sequences to one another. Unlike the prior metrics, it does not use a precise formula, but can be solved algorithmically:

1. Begin with sequences T_{chr} and T_{txt} and $d_C = 0$.
2. Compare $T_{chr}(0)$ and $T_{txt}(0)$.
    - If $T_{chr}(0) \neq T_{txt}(0)$, add one to d_C and locate i such that $T_{txt}(i) = T_{chr}(0)$. Modify T_{txt} by swapping $T_{txt}(0)$ and $T_{txt}(i)$.

[59]See Irurozki et al. (2016).

3. Take the current version of T_{chr} and T_{txt} and drop the first item from each sequence. Every index for each remaining item is decremented by 1.

4. If $Len(T_{chr}) > 1$, return to Step 2. Repeat until sequences are length 1.

This value is upper bounded by (but not exactly the same as) the *number of items in the final sequence which are not in the same position as the original sequence*[60]. To normalize Cayley distance, divide by its maximum possible value, which is $n - 1$ for a sequence of length n.[61].

In Table 6, these three metrics are applied to timeline types from the earlier taxonomy.

Timeline Type	d_S	d_K	d_C
Chronological	0	0	0
Mostly chronological	.16	.1	.25
Inverted	**1**	**1**	.667
Disordered	.8125	.536	**.714**

Table 6: Normalized metric per sample timeline-pairs.

Across all selected metrics, the chronological timeline has a score of 0, and all others have values greater than 0. The mostly-chronological timeline has low scores, and the inverted/disordered timelines score higher across all metrics. The two types of timelines which feature the most temporal deviation (inverted and disordered) score differently across metrics. For two (Spearman's and Kendall), the inverted timelines scores having the most temporal deviation, while Cayley's placed the disordered timeline as most deviant.

Overall, these metrics provide useful tools for examining timelines across IlNeS, though they present only broad understandings of timelines and distinct timeline types. To truly understand the mechanisms behind timeline deviations, the dissertation proposes a paradigm of timeline interpretation which provides deeper insight into how individual timelines function.

5.3.3 Timeline Leaf Paradigm

The timeline leaf paradigm asserts that temporal deviations in timelines often follow regular patterns, and distinct segments (or 'leaves') within the same timeline may behave as different timeline types. Consider another example (Figure 50) of a disordered timeline taken from IlNeS.

	Timeline 1334
d_S	.264
d_K	.145
d_C	.6

Table 7: Metric values for Timeline 1334.

The distance scores (Table 7) for this timeline type position it as more deviant than the mostly-chronological timeline type, but less so than the inverted or the other disordered example. This timeline sample is highly chaotic to a human reader, but it can be re-framed as a collection of smaller timeline segments (called "leaf timelines").

[60]If two items are in different positions from in the original sequence, but can both be placed in order by a single transposition, the contribution to the final Cayley value is 1 and not 2.

[61]See Irurozki et al. (2016).

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"done everything"	"unremitting sadness"	"broadsided"	"contain my sadness"	"dealt with bipolar"
Textual:	"done everything"	"broadsided"	"unremitting sadness"	"will myself to"	"look for opportunities"
	Position 6	**Position 7**	**Position 8**	**Position 9**	**Position 10**
Chronological:	"will myself to"	"look for opportunities"	"allowed me"	"slip home"	"guarantee"
Textual:	"slip home"	"contain my sadness"	"dealt with bipolar"	"allowed me"	"guarantee"
	Position 11				
Chronological:	"find footing"				
Textual:	"find footing"				

Figure 50: Disordered timeline example, from Participant 1334.

Figure 51 (see next page) splits Timeline 1334 into two "leaves", demonstrating that Leaves 1 and 2 fall under the **mostly chronological** category–that is, for almost all pairs of events within Leaf 1, the ordering is chronological. The majority of temporal deviations in the full textual timeline come from *inter-leaf interactions*. Further, events in leaves correlate with distinct topics in the text itself: the events in Leaf 1 describe the **symptoms** of the participant's condition, and Leaf 2 the **recovery**. This has implications for how human narrators may mentally organize narratives before arranging them in text, and suggests semantic similarity could be provide useful features to a predictive model.

Additional examples of multi-leaf timelines can be found in the Appendix in 10.4, with explanation of how semantic understanding of each text informs segmentation of each leaf.

Surfacing Leaves:

Leaves and their interactions can be surfaced by counting "steps" in the textual sequence. In ranking metric calculations, the value of any $T_{txt}(i)$ is equivalent to its ranking inn T_{chr}. For each position i, the step value is as follows:

$$Step(i) = T_{txt}(i) - T_{txt}(i-1) \tag{7}$$

If $i = 0$, the equation is instead:

$$Step(0) = T_{txt}(0) \tag{8}$$

In a truly chronological sequence, the value of $Step(i)$ will be 0 (representing events which are simultaneous and share a chronological rank) or 1 (representing a single step forward in the timeline) for all i. Minor temporal deviations will manifest as slight dips and peaks in the step graph. Larger discrepancies mark spots where the timeline jumps between leaves–surfacing these distinct elements of the timeline. An example is shown in Figure 52 with the step visualization of Timeline 1334. On the graph, the notable steps match the switches between leaves which occur at positions 4, 7, and 9.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"done everything"	"unremitting sadness"	"broadsided"	"contain my sadness"	"dealt with bipolar"
Textual:	"done everything"	"broadsided"	"unremitting sadness"	"will myself to"	"look for opportunities"
	Position 6	**Position 7**	**Position 8**	**Position 9**	**Position 10**
Chronological:	"will myself to"	"look for opportunities"	"allowed me"	"slip home"	"guarantee"
Textual:	"slip home"	"contain my sadness"	"dealt with bipolar"	"allowed me"	"guarantee"
	Position 11				
Chronological:	"find footing"				
Textual:	"find footing"				

Leaf 1	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	
Chronological:	"done everything"	"unremitting sadness"	"broadsided"	"contain my sadness"	"dealt with bipolar"	
Textual:	"done everything"	"broadsided"	"unremitting sadness"	"contain my sadness"	"dealt with bipolar"	
Leaf 2	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**Position 6**
Chronological:	"will myself to"	"look for opportunities"	"allowed me"	"slip home"	"guarantee"	"find footing"
Textual:	"will myself to"	"look for opportunities"	"slip home"	"allowed me"	"guarantee"	"find footing"

Figure 51: Timeline 1334 split by leaf to identify specific behaviors within timeline.

(Further examples may be found in the Appendix 10.4.)

This approach allows timeline analysis to be more precise, and proposes the following: that there are two types of temporal deviations that may be found from T_{chr} to T_{txt}. First is **local perturbations**, which are most commonly attributable to grammatical and semantic tendencies inherent to the text's language. The following are proposed as examples of possible behaviors which may form local perturbations (note that this dissertation examines the behavior of English language, and may not generalize to languages beyond):

- Cause-effect switching: "A happened because B", where causal event B is presented after the effect A it caused. Likely to explain a switch between two neighboring events from T_{chr} to T_{txt}.
- Topicalization: The movement of the central focus of a text to the beginning of that text. Likely to explain a relocation of some event in the middle of T_{chr} to the start of T_{txt} (or, in the case of a multipart timeline, a smaller segment).

Local perturbations determine whether a timeline qualifies as chronological or mostly chronological (if the timeline is inverted, local perturbations may also be applied to produce a mostly inverted) timeline. These are considered to be the three fundamental types of timeline units, as disordered timelines are typically better described as 'multi-leaf'. Leaves within a disordered timeline can be defined using the 3 fundamental sub-types.

Segment perturbation occurs in two types: 1) inversion of a segment; 2) inter-leaf inter-

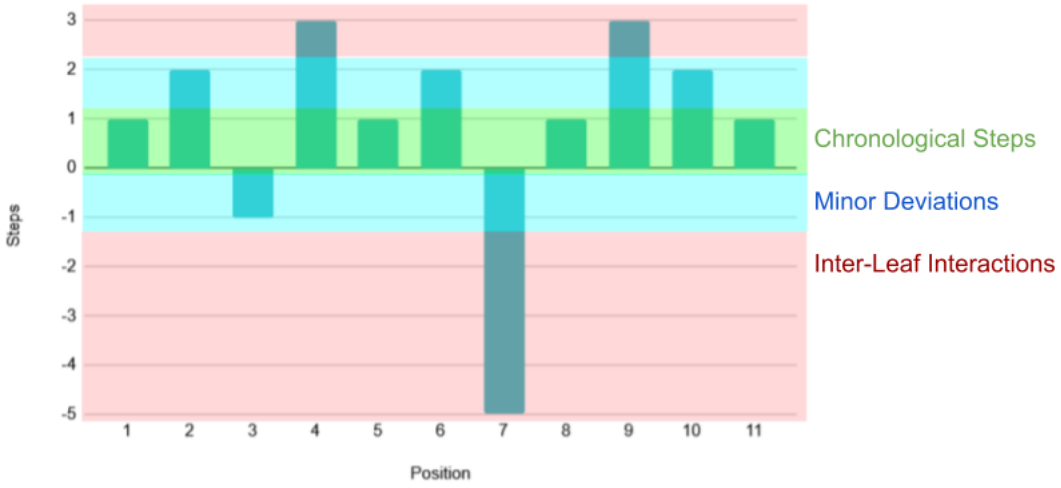

Figure 52: Step visualization of Timeline 1334, with distinct types of deviation marked.

actions which occur between segments. Leaves of a timeline may be arranged in **disjoint** segments, where where there is notable overlap across leaves (see Timeline 6282, Figure 76 in Appendix 10.2.2) or in **discrete** segments, where leaves do not or mostly do not overlap (see Timeline 7542, in Figure 78 of Appendix 10.2.2).

Segment-level perturbations suggest a deeper motivation than local deviations. Non-fiction writing guides recommend that text be organized by *concept* rather than by *time*; this approach often matches a chronological timeline, but ultimately supersedes it. "[The conceptual approach] works for all kinds of stories; it is almost infinitely flexible. Even stories that seem to be put together chronologically are in fact organized conceptually" (Fontaine et al., 1987). In the Appendix, Section 10.2.2, the dissertation shows multiple examples that each leaf in a timeline can be tied together with a unifying semantic concept within the participants' experience. This suggests this rule of nonfiction is understood and repeated, even if only unconsciously, by lay writers across the survey demographic.

5.3.4 Label Distribution

Metrics such as d_S, d_K, and d_C all effect the distribution of labels within the final MulCo input. This responds to the dissertation research question: **Are there certain types of narrative which exhibit temporal deviations disproportionate to prior class biases of ordering models?** A challenge in TEO/ETRE is that corpora often have strong biases towards a specific relation label (typically $before$ or $vague$). Models learn these biases and may produce high F1 scores while behaving very similarly to a simple majority classifier. If a model does not differ from the behavior of a majority classifier, its architecture cannot demonstrate anything meaningful about the dataset or the task itself.

Analysis of labels within a corpus asks the following sub-questions:

- Do the distributions of temporal deviations in different narrative types introduce significant bias towards the majority classifier?
- Can the IlNeS dataset be adjusted to ensure models learn more than simple majority

classification?

5.4 Results

Results of evaluation methods are discussed here.

5.4.1 Metrics

In Table 8, the average values for each ranking metric are displayed for each NarraType subset in IlNeS.

	d_S	d_K	d_C
Factual	.121	.083	.218
Persuasive	.244	.187	.392
Emotional	.247	.164	.398
Clinician	.200	.148	.327

Table 8: Average normalized metric for each NarraType in IlNeS.

To identify if these differing averages reflect statistical significance, the work evaluates the distribution of metrics per document. Comparison of NarraType ranking distributions (Table 9) does show significant results for both Spearman's footrule and Kendall tau distance (within $p < .05$). This comparison is performed using two-sample Kolmogorov-Smirnov fit test.

p-value	d_S	d_K	d_C
Factual + Persuasive	**.047**	**.047**	.283
Factual + Emotional	**.045**	**.041**	.197
Factual + Clinician	.853	.365	.853
Persuasive + Emotional	.785	.773	.996
Persuasive + Clinician	.410	.410	.768
Emotional + Clinician	.632	.387	.833

Table 9: Significance scores, with $p < .05$ bolded.

Two-sample KS tests show significant differences between the Factual and Persuasive/Emotional narratives. Note that, while the differences between Factual and Clinical scores near those of Factual + Persuasive/Emotional, they do not fall within statistical significance. There is no observed statistical significance between any two NarraTypes for Cayley's distance. Overall, the Factual NarraType shows the closest adherence to chronological timelines, followed by Clinician. There is little observed difference between Persuasive and Emotional NarraType.

Examining the distributions of metrics for each NarraType's timelines shows interesting trends when graphed against the length of the timeline (see Figures 53, 54, 55).

These graphs demonstrate:

1. Temporal deviations are most noticeable when size of timeline n is smallest across all NarraTypes. However, variability in these smaller timelines mean the average remains steady.
2. Factual timelines have the least variation in behavior based on the size of the timeline compared to other NarraTypes.

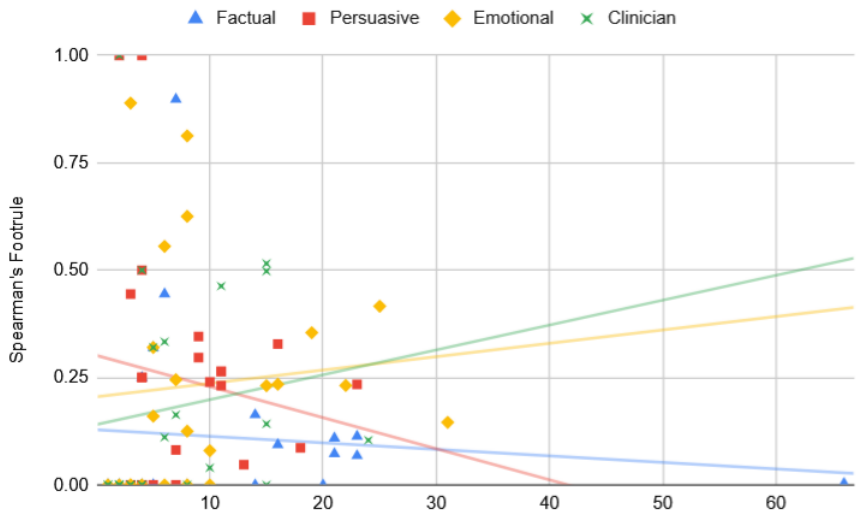

Figure 53: Distribution of Spearman's footrule (y-axis) compared to timeline size (x-axis) across IlNeS.

3. Persuasive timelines sharply decrease for certain ranking scores as timelines get larger.
4. Emotional and Clinician timelines see increases in all ranking metrics as timelines get larger, and behave almost identically by all three measures.
5. Clinician and Factual timelines show the most variability which cannot be explained by the size of timeline alone. For Clinician timelines, this could be explained as the pseudo-NarraType behaving as a mixture of multiple prompts.
6. Factual timelines are most influenced by notable and specific outliers that fell outside usual trends. These outliers could not be removed from the data without justification, but have a disproportionate mathematical impact on trend-line calculation.

These results give early suggestions that Clinician narratives, rather than matching any one of the original three NarraTypes, vary based on how participants prioritize their needs (though it behaves most like the Emotional NarraType). Factual timelines are more ordered and chronological regardless of timeline size, and Persuasive timelines show distinct behavior compared to other NarraTypes.

5.4.2 Labels

IlNeS differs from previous TEO/ETRE corpora in that it *comprehensively* covers all pairs per document. In the four datasets which annotate for TimeBank documents (TimeBankDense, TDDiscourse-Manual and -Auto, MATRES), annotations cover only a subset of the full pairings possible. To extract label distributions across the entire document set, multiple corpora must be combined.

TimeBankDense, TDDiscourse-Manual, and TDDiscourse-Auto represent mutually-exclusive subsets of the TimeBank text corpus[62] Because the design of TDDiscourse allowed duplicate

[62]MATRES, which has significant overlap with TimeBankDense (as reported in Ning et al., 2018), was omitted

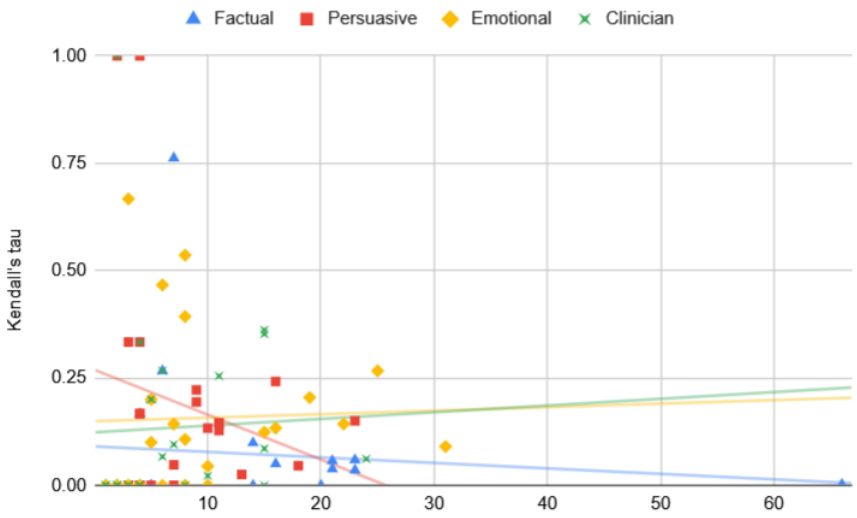

Figure 54: Distribution of Kendall's tau distance (y-axis) compared to timeline size (x-axis) across IlNeS.

	Bef	**Aft**	**Sim**	**Inc**	**IsInc**
IlNeS	.75	.08	.05	.10	.02
TimeBank (agg.)	.32	.27	.14	.14	.14

Table 10: Proportion of gold-standard labels in IlNeS compared to aggregated TimeBank.

annotation of a given event-pair when the original TimeBankDense annotation was *vague*, these pairs have also been omitted. The final counts use only non-*vague* event-pairs across these three corpora to produce aggregated label distributions across TimeBank, shown in Table 10.

The distribution of labels in IlNeS differs significantly from the available labels for TimeBank documents. Note that TimeBank has not been annotated comprehensively. It is possible that pairs which remain un-annotated in the TimeBank corpora would lead to total distributions which match IlNeS. As current models are trained on corpora derived from TimeBank, their architecture may not compensate for class biases required to predict for IlNeS.

5.5 Analysis for Confounds

Two potential confounding variables exists in the data: 1) the **clarity of intent** of the participant in a given prompt, and 2) the **illness type** which forms the topic of the testimonial. In this section, these confounding factors are examined to determine what, if any, impact they have on the relative disorder of a given timeline. This section also addresses individual testimonials which have a disproportionate impact on statistical results; though these outliers cannot be excluded from analysis without justification, the effects they have on the data is worth addressing and acknowledging.

for this analysis.

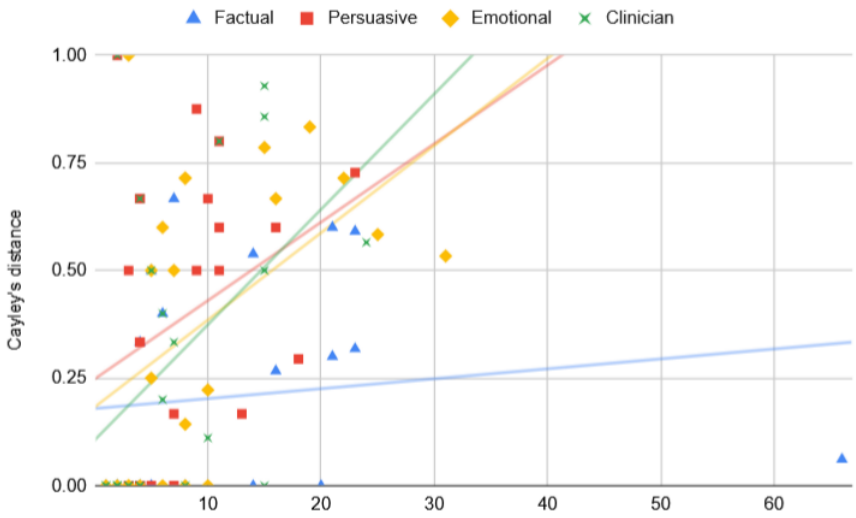

Figure 55: Distribution of Cayley's distance (y-axis) compared to timeline size (x-axis) across IlNeS.

5.5.1 Clarity of Intent

The proxy audience prompt was decided to elicit specific authorial intentions from participants without directly asking them to write in a specific way. This requires that participants understand and engage with the audience prompt, and that the IlNeS prompts are successful in isolating the mechanism through which authors adapt texts to audience (or "intention"). This segment of the work examines a potential confound in IlNeS analysis: that temporal disorder in a text may instead correlate with clear expressions of intent (i.e. orientation towards the desired IlNeS response) rather than the type of intention itself.

This variable–which will be called **intention clarity** in this section–is captured as a Boolean value rather than a sliding scale, due to its potential for subjectivity[63]. It is judged based on statements which 'frame' the text as having a specific purpose, or statements from the participant's freeform feedback with express an engagement with the underlying desired NarraType[64].

All participants who provided a testimonial engaged with the prompt in *some* way; the variable of clarity separates texts where participants are addressing some ambiguous audience and where participants can be stated definitively to have shaped their text with intention. This analysis uses specific criteria per prompt, in order to ensure consistency:

- Criteria for Factual NarraType: references are made (in either testimonial or feedback) to 'informing', 'explaining', or 'conveying' certain information to the audience.
- Criteria for Persuasive: direct references are made to 'persuasion' (or synonyms); or the testimonial/feedback articulates distinct purposes the imaginary funds or research resources discussed in the prompt could be used for.

[63]Though annotations of engagement level along some scale (ex. Likert) presents interesting avenues for future study.

[64]Examples of these discussed in Section 6.1 as part of IlNeS design validation.

- Criteria for Emotional: expressions of sympathy, support, or emotionally-centered advice for the imagined audience.

Because the Clinical NarraType does not encode a distinct intention (and is used to judge which intentions patients bring to clinical interactions), a Clinical testimonial is judged as having clear "intent" if a direct reference is made to the imaginary GP in the testimony or feedback.

As the goal of this analysis is to test the effect of intention clarity on temporal disorder within a narrative, it uses the subset of IlNeS for which participants annotated a timeline with at least one event-pair (i.e. where length $n \geq 2$). This reduces the number of analyzed texts from 106 to 89. There are individual cases in this remaining corpus where a participant shows engagement with a hypothetical audience but does not meet the criteria for specific orientation to the prompt. For example, Participant 2492 (who was assigned the Persuasive NarraType) states in feedback, "Was trying to recall all important information about the event without disclosing personal details and give them in a concise and clear manner, while retaining chronological order." This statement does not indicate that the details about disclosure and order are done to achieve a specific persuasive goal, and so does not meet criteria for intention clarity. 4677 (assigned the Factual NarraType) states, "I was just thinking about my own personal experience with depression as it is a disability for me, and thinking about how some people may not take it seriously or even believe that it is a disability." There is no mention of informing, explaining, or even clarifying the basic condition to an acquaintance as specified in the prompt.

Variable Evaluation:

Though this variable type was suggested as a potential confound for statistical analysis, it is not necessarily independent of the assigned NarraType variable. Across the corpus, the chosen criteria produce distinct distributions per NarraType, shown in Table 11.

NarraType	Intention Clarity Count	Total Count	Proportion
Factual	10	21	.476
Persuasive	18	24	.750
Emotional	22	22	1.00
Clinician	11	22	.500
Total	61	89	.685

Table 11: Counts and proportions of testimonials per NarraType such that participants show clear engagement with the intention task.

There is observable variance in the proportion of participants engaged with the task based on NarraType. This could reflect some inherent property of the NarraTypes in question, and in turn the behavior of human authors in natural text, or it could be a consequence of the IlNeS study design. The two partitions showing lowest levels of clarity are Factual and Clinician; the Clinician NarraType explicitly does not direct participants to a specific intention, and Factual may suffer from being perceived as the most 'neutral' mode of authorial intent. It is possible that high-clarity Factual testimonies can be collected using a different audience proxy prompt, or it may be that Factual prompts will always induce less intense feelings of intention from an audience compared to Persuasive or Emotional.

This section tests the significance of the observed variance by comparing against a null hypothesis that intention clarity will follow a binomial distribution that holds for all NarraTypes. To use a visual example, if the variable of intention clarity follows a binomial distribution, then the probability of any distribution of size $n = 22$ having x total successes (i.e. the **probability**

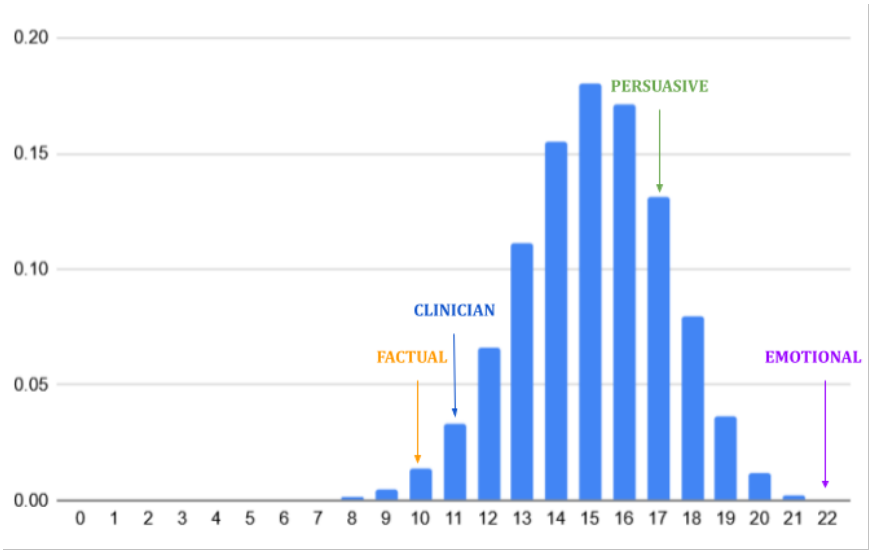

Figure 56: A demonstration of $P(X = x)$ for distribution size $n = 22$ if the odds of positive intention clarity are random and independent of NarraType. Markers are placed to approximate the odds of seeing the partition's distribution in the hull hypothesis.

distribution function) will roughly[65] match the graph shown in Figure 56. To test the validity of the null hypothesis, this analysis uses the **cumulative distribution function** where $p_{success} = .685$ for all distributions, and where number of successes x and distribution size n depend on observed results for each IlNeS partition.

When testing the probability of a set of successes against a cumulative distribution function, the p-value should sum the probability of all distributions of successes which are *as extreme* or *more extreme* compared to the observed output. Therefore, each NarraType has an associated $P(X \leq x)$ or $P(X \geq x)$ used to calculate the final p-value. (Note that for the Emotional NarraType, it is not possible to produce more than 22 successes for a distribution of $n = 22$, so the comparative probability is $P(X = 22)$.)

The full results of this comparison are shown in Table 12.

NarraType	Distribution Tested	Distribution Size	p-value
Factual	$P(X \leq 10)$	21	**.038**
Persuasive	$P(X \geq 18)$	24	.329
Emotional	$P(X = 22)$	22	**2.4e-4**
Clinician	$P(X \leq 11)$	22	.054

Table 12: The p-value of the null hypothesis that the distributions of intention clarity in IlNeS follow a binomial distribution.

Based on this output, it is possible for both the Persuasive and Clinician NarraType distributions of intention clarity to be the result of random chance compared to the binomial

[65]Figure 56 uses $n = 22$ for all partitions to simplify the visual. The markers in the image shift outputs for partitions of different size to their nearest proportional number of successes, but do not reflect an exact comparison.

distribution. Factual and Emotional demonstrate similarity with $p < .05$ when compared to this same distribution. This suggests that these variables are not fully independent of one another. Intention clarity was introduced as a possible confound to the findings of this section, but may instead illuminate a specific mechanism through which temporal disorder is introduced to testimonials.

Impact: As was done to test the three 'disorder' metrics per NarraType, two-sample Kolmogorov–Smirnov testing is done to isolate the effect of the intention clarity variable. For each NarraType (and for the total distribution), data is split into sets such that $IntClar(Participant) = True$ for all participants in one set and $IntClar(Participant) = False$ for the other. The metric compared for this testing is Kendall's distance. Two-sample KS testing identifies the probability against the null hypothesis that both sets come from the same underlying distribution.

(Note that the Emotional NarraType partition cannot be tested for this variable, as there are 0 total Emotional NarraType testimonials which did not meet the criteria for intention clarity.) Results are shown in Table 13.

NarraType	p-value
Factual	.227
Persuasive	.302
Emotional	N/A
Clinician	.833
TOTAL	**.022**

Table 13: The p-value of the null hypothesis that the distributions of intention clarity in IlNeS follow a binomial distribution.

No NarraType distribution in isolation shows statistically significant impact for intention clarity. The distributions across all NarraTypes do show statistical significance for $p < .05$ (similar to the significance level for differences across NarraType). However, this specific result does not account for the impact of NarraType on intention clarity. This suggests two possible explanations for the data shown: 1) that intention clarity has some correlation with temporal disorder across narratives, but this correlation is itself related to the impact of NarraType on temporal disorder; 2) that intention clarity is a confounding variable for these metrics, but that this effect requires larger datasets the individual IlNeS partitions to surface.

5.5.2 Illness Type

Another potential confound is the type of illness experienced by participants. Arthur Frank's conception of the illness narrative genre positioned it as responsive to an author's feelings about their condition. Different types of illness may evoke distinct responses, which may in turn influence the temporal order of events within the narrative independent of NarraType.

Because the research team had no knowledge of participants' medical conditions before assigning them a random NarraType prompt, there is a strong case to be made that illness type is *mostly independent* of this variable. However, it may not be fully independent of NarraType: a participant from the IlNeS pilot phase reported[66] that they had multiple medical conditions which could have been used as the topic of the testimonial (call these Conditions A and B), but they chose to write about Condition B because they had been assigned the Persuasive

[66]Specific feedback repeated in this dissertation with permission.

prompt. This participant's reasoning was that they felt Condition A was well-funded and well-researched, while Condition B could use additional resources.

This introduces the possibility of minor dependence between variables, in the specific circumstance where a participant has multiple medical conditions they might choose from. The work theorizes that this dependence might arise for the following reasons:

- A participant assigned the Factual NarraType, told to imagine their audience does not know much about their condition, may find the prompt more believable if the chosen medical condition is not well-known.
- A participant assigned the Persuasive NarraType may choose the condition of theirs for which they feel additional funding and resources are most necessary.
- A participant assigned the Emotional NarraType may choose the condition which the greatest emotional impact on them personally.
- Management of trauma–there may be some imagined audiences for which speaking about a specific illness might feel traumatic or inappropriate.

Information provided about participant medical conditions in IlNeS was quite diverse; see Table 41 in Appendix 10.3.1 for a full list of conditions represented in the corpus. However, the specificity that was often provided hinders statistical analysis, as many conditions appear infrequently. This analysis clusters medical conditions into 7 categories (see Table 14), each covering at least 10 testimonials. Note that all medical conditions are provided and described by participants. A condition which a given participant considers a 'disability' may not be considered as such by all individuals with this condition.

Illness Cluster	Count	Examples
Potential threat to life	12	Cancers, cardiovascular conditions, COPD, liver failure
Autoimmune/immune disorders	10	Asthma, celiac, mast cell, psoriasis
Chronic (hormonal)	10	Diabetes, endometriosis, PSSD
Chronic (other)	11	Epilepsy, fatigue, idiopathic conditions
Co-morbid or Unspecified	15	Participants discussed multiple conditions equally, or did not provide an exact condition
Mental disability	15	Anxiety, depression, PTSD, schizophrenia
Physical disability	16	Chronic pain conditions, loss of mobility
Total	89	

Table 14: Clusters of illness types present in the IlNeS corpus, with counts and examples of illness falling in each cluster.

An effort was made to group these illnesses not only by direct medical similarity, but by impact on participant. A chronic condition which is not life-threatening but requires consistent management may be perceived differently to an acute and potentially terminal condition, for example. Some edge cases required careful individual consideration to determine a best "fit". Many conditions have potential to be life-threatening (ex. epilepsy) or do not have exact known cause (endometriosis), but in practice behave closer to chronic conditions grouped by some other defining factor.

Because these clusters are small ($n \leq 16$), it is not possible to draw statistical conclusions about their distribution across NarraType partitions except for extreme causes. The full distribution

can be found in Table 15.

Illness Cluster	Fact.	Pers.	Emot.	Clin.	Total	Intent Clarity
Potential threat to life	1	4	3	4	12	.75
Autoimmune/immune disorders	2	2	2	4	10	.5
Chronic (hormonal)	2	4	1	3	10	.8
Chronic (other)	5	4	2	0	11	.727
Co-morbid or Unspecified	2	4	6	3	15	.733
Mental disability	6	5	3	1	15	.667
Physical disability	3	1	5	7	16	.625

Table 15: Distribution of NarraType per illness cluster.

Notable observed differences include the distinct distributions across the **physical disability** and **mental disability** categories. In studying distributions of these clusters, physical disability had few examples of Factual or Persuasive narratives, but many Emotional or Clinician narratives. Mental disability showed the reverse; it is possible that physical conditions are considered to be less in need of education or funding compared to mental health conditions, but still require significant emotional support. Culturally in the United States (where this survey was collected), mental health conditions have historically been stigmatized, and the shift in approval surrounding mental heath care is relatively recent. Therefore, participants may see more need to educate their acquaintances or make pleas for more research and funding for these conditions compared to physical conditions. The infrequency of Clinician texts for this category could be due to an unwillingness to bring up mental health to a medical provider, or general practitioners being considered less connected to mental health care.

A similar noteworthy difference can be found in the **chronic (other)** conditions, where almost half of participants with these conditions provided Factual narratives and none provided Clinician. If there is dependence between this cluster and NarraType, it may relate to the presence of *idiopathic* conditions within the cluster. Idiopathic conditions (i.e. a medical condition with no clear cause) create unique burdens for patients; patients are forced to pay close attention to their daily lives for signs of potential triggers and treatments. This type of self-expertise may be dismissed by clinicians, and when a condition is rare they may refuse to believe the patient is experiencing it. Even when clinicians are supportive, idiopathic illnesses often suggest a prolonged experience with the medical system as providers seek to diagnose the condition. Therefore, IlNeS participants might have aversions towards discussing these condition types with clinicians, but less hesitance about doing so with lay acquaintances.

It is important to remember that the **medical conditions experienced by the participant** and **the NarraType prompt the participant is assigned** are independent variables from one another. Participants enter the study with a fixed medical history, and the survey assigns NarraType at random knowing nothing of that history. The illness type cluster a participant's testimony belongs to depends on the **illness that is chosen as the survey topic**. If this variable is dependent on the previous two, it is an *interaction* between these variables which informs a participant's choice from within their medical history (visualized by Figure 57).

Impact: No pairing of illness cluster distributions shows statistical significant differences for Kendall's distance. Like the prior analyses, this evaluation used 2-sample KS testing with p-value $p < .05$ to compare the distributions of temporal disorder per illness type. The full results are shown in Table 16.

Though the illness type clusters were arranged to provide decent of n, a larger dataset sorted

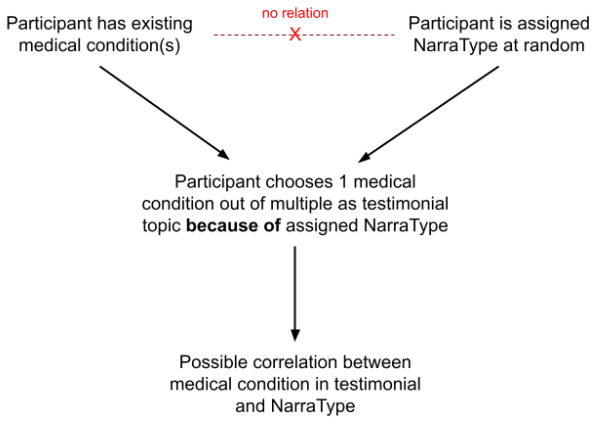

Figure 57: Visualization of the process by which the illness chosen in testimonial may correlate with NarraType despite the random assignment of NarraType prompts.

by illness type could provide more definitive insight on whether illness type impacts temporal disorder. 'Illness type' is also a challenging category to break into discrete clusters, as medical conditions can be compared based on the bodily system they effect, underlying cause, treatment type, long-term impact, or even level of function[67]. This particular set of clusters was intended to maximize the size of similar groupings across IlNeS participant testimonials, but may not reflect the best organization of conditions supplied by the corpus.

5.6 Discussion

To answer RQ1. **How does the purpose of a narrative change the order in which temporal events are presented?**, the work does see differences in the temporal behavior of text within IlNeS testimonials based on NarraType. Factual NarraType testimonials have the least amount of temporal deviation, and have statistically-significant differences in rank similarity scores compared to Persuasive and Emotional. Persuasive and Emotional demonstrate extremely similar scores for all categories, even when graphed based on document length. Clinician NarraTypes do not significantly differ from any category.

This supports an early hypothesis of the work–that Factual narratives would most closely match chronological ordering compared to other NarraTypes–but is not alone proof of that hypothesis. Observing the distribution of each metric per NarraType, charted against length, does show that Factual testimonials maintain a more consistent number of deviations even as the size of the document grows than other NarraTypes.

RQ2. **What can we observe about the link between temporal positioning in text and intended effect from short-form standardized narrative text?** is answered by a set of rank metrics the work adapted for timeline sequence analysis, typologies of timeline behavior which include leaves behaving as sub-timelines in a full text, and step discrepancy methods which can surface points of significant segment-level temporal deviation. These methods allow

[67]For example, WHO, 2022 and WHO, 2001 include coding systems to describe disease, disability, and daily function which are useful for clinical annotation but in this case would be too granular.

Cluster Comparison	p-value
Potential threat to life + Chronic (hormonal)	.996
Autoimmune/immune disorders + Chronic (hormonal)	.994
Potential threat to life + Autoimmune/immune disorders	.953
Autoimmune/immune disorders + Physical disability	.923
Chronic (hormonal) + Physical disability	.651
Chronic (hormonal) + Mental disability	.333
Chronic (other) + Mental disability	.333
Chronic (other) + Physical disability	.29
Potential threat to life + Mental disability	.267
Chronic (other) + Co-morbid or Unspecified	.267
Autoimmune/immune disorders + Co-morbid or Unspecified	.233
Potential threat to life + Physical disability	.229
Co-morbid or Unspecified + Physical disability	.221
Chronic (hormonal) + Chronic (other)	.209
Chronic (other) + Co-morbid or Unspecified	.194
Mental disability + Physical disability	.183
Potential threat to life + Chronic (other)	.167
Potential threat to life + Co-morbid or Unspecified	.167
Autoimmune/immune disorders + Mental disability	.167
Autoimmune/immune disorders + Chronic (other)	.164
Chronic (hormonal) + Co-morbid or Unspecified	.133

Table 16: p-values for 2-sample KS tests across all illness type combinations.

for a wide range of analyses to be applied to timelines, either in deep study of an individual text or across a corpus.

RQ3. **Are there certain types of narrative which exhibit temporal deviations disproportionate to prior class biases of ordering models?** This research question is not fully answered by this chapter, but the rank metric work suggests trends which will be examined further in Chapter 8. Based on this section's output, it would be expected that Factual narratives will contain fewer temporal deviation labels ($after$ and $isincluded$) compared to other NarraTypes, and that Persuasive and Emotional may have similar distributions of TEO/ETRE labels overall.

An additional sub-question driving the research asks if the behavior of participants given the Clinician prompt would match that of any other NarraType (as this might inform patient-clinician communication). The positioning of Clinician prompts in the middle of other NarraTypes suggests that Clinician testimonials are highly individual and may draw from any of our three main intention types.

5.7 Contributions

This dissertation contributes a complete, annotated corpus, IlNeS, for the illness narrative domain. It provides the following data for future use in other studies:

1. 106 unique illness narratives, collected from participants who have experienced chronic or acute medical conditions. These narratives were elicited in rich text format using prompts for imagined audience types.

2. Annotations for the prompted audience type per narrative.
3. Annotations for relevant "medical events", as noted by participant.
4. Annotations for the temporal order of all medical events in text, as noted by participant.
5. Freeform text responses describing the process of writing the original illness narrative. Though the quality of these responses varies per participant, many provide rich insight into the process of recalling events and developing narrative.

Further, this dissertation presents layered analysis of the IlNeS corpus. Through IlNeS, it is possible to more rigorously examine existing theories of the illness narrative field. This dataset and all annotations will be released in anonymized format for public use after dissertation completion.

6 IlNeS Testimonial Analysis

Though IlNeS was collected for the primary purpose of validating *temporal hypotheses* about time, there is rich data to be gathered about the experience of illness and the genre of illness narrative using the testimonials themselves. The framework of narrative formalized by Genette (1980) includes three primary units: the content (or 'story'), the delivery ('narrating'), and packaging ('discourse'). In the IlNeS corpus, the content of a participant's experience is as static as can be reasonably achieved, while the method of narrating is influenced by the proxy prompts simulating distinct audiences. Therefore, analysis of illness testimonials within IlNeS allows the dissertation to isolate and evaluate the way that the *narrating* unit of text in Gerard's framework impacts the *discourse*.

This chapter will analyze the provided testimonials using close and deep reading techniques, to answer Research Question 4: **How is the behavior of short-form illness narratives driven by the factor of authorial intent?**

6.1 Proxy Prompt Validation

The survey design (as discussed in Chapter 5) used a proxy variable of "imagined audience" to capture NarraType; the collected testimonials and freeform feedback question validate that the survey design achieves this goal. This section steps through each NarraType and provides evidence that participants assigned the corresponding prompt were aligning their intention with the NarraType. Sample text is collected both from participant testimonials and from the freeform Question 4 which allowed participants to describe their thought processes to the researchers. All examples which validate the survey design for each partition can be found in the appendix, under Section 10.3.2.

6.1.1 Factual Scenario

The Factual scenario presented participants with a hypothetical audience who did not know basic details of the participant's condition. This scenario was designed to prime participants towards the goal of using the testimonial narrative to educate their audience.

In collected testimonials, framing statements from within the text that participants assigned this prompt had made an effort to write as if communicating with someone lacking education in their condition. For example:

> "Hey, I have never really explained this to you before"
>
> -Participant 0775, within the illness narrative

Freeform feedback responses also support this assertion. To quote another participant:

> "I tried to explain my depression as if I was explaining it to someone who has never had experience with it [...] I wanted to try to put things in perspective and simplify as best as I could for a maximum understanding of a complex issue."
>
> -2609, in freeform feedback

Overall, participants given the Factual prompt were clearer about their intentions within the freeform feedback section than in the narrative itself. They often referenced (in freeform feedback) a desire to "explain" their condition. Many participants within this partition cited aims

to be "honest" and "clear"[68]. Full compilations of Factual proxy design validation (located in Appendix 10.3) are displayed in Tables 43 and 44.

6.1.2 Persuasive Scenario

The Persuasive scenario presented participants with an imagined audience who could provide them with something they may desire (e.g. improved research or treatment for their condition). This was designed to encourage participants to approach their narrative as an opportunity to persuade the listener to help their own goals.

Select responses from participants are shown here. The full collection can be found in the Appendix 10.3 in Tables 45 and 46.

> "Hi, I write to you in times of desperation and just wanting to live a normal life. [...] I urge you politely and respectfully to allocate funds for better treatment and accommodations for individuals like me so that we may have a chance of living a somewhat normal and quality life."
>
> -1464, in illness narrative

An example from feedback:

> "I was thinking about how to balance personal experience with persuasive impact. [...] I wanted the tone to be respectful but urgent, and I aimed to emphasize both the human and economic cost of underfunding treatment and accommodations."
>
> -2111, in feedback

The tone of Persuasive narratives was often respectful and showed deference to the imagined audience. There were cases of emotional wording within select texts[69]. A large proportion of narratives textually addressed the imagined audience. In feedback statements, participants directly articulated a desire to persuade; some contained detailed description of how the participant believed persuasion could best be achieved. Some participants noted the economic element of disability, and others the personal impact. Despite this, many participants noted a desire not to appear emotional (to be discussed more in Section 6.3.4).

Time is a focus of the dissertation. It is therefore notable that in one case, a participant expressed direct intention about temporal deviation within the text: Participant 2492 used the feedback form to state their intention had been to produce a fully-chronological timeline in their narrative text.

6.1.3 Emotional Scenario

The Emotional scenario was designed to inspire empathy among participants and a desire to forge an emotional connection with their imagined audience. One expectation of the Emotional scenario was that participants might use the narrative to give advice that they wish they had been given at the start of their own illness journey. In this way, participants would imagine themselves in place of their audience, in keeping with insights from Frank and Bury about the purpose of emotional illness narratives.

[68] Though both aims were also present to a lesser degree in feedback for other NarraType scenarios.

[69] While neutral tones may be understood to be more effective in persuading an audience, the imagined scenario of a researcher or policymaker potentially *withholding* necessary resources inspired anger in some participants.

Narratives within the Emotional NarraType partition had more variety than other NarraTypes. This suggests that emotional intent is most sensitive to the mindset of the speaker. Many participants provided their imagined audiences with practical, medical, or even vocational advice from the perspective of one who has been in such a position before. Though this may seem to contradict the idea that the proxy audience prompt elicits emotional behavior from participants, it is similar to Rini et al. (2014)'s framework of *expressive helping*, where advice and encouragement (peer-helping) combined with opportunities to express emotion through writing (emotional-writing) reduced distress among cancer survivors within a cohort. In at least some cases[70], helping others can allow oneself to feel happier and more fulfilled.

In a turbulent period of life (such as during illness), acting as a confident expert may help participants reassert a sense of control. Testimonials which are pragmatic and direct therefore do not contradict the assertion that the primary intention behind the text is emotional.

The full selection of quotes can be found in Tables 47 and 48 within Section 10.3.

> "I hate to see that you're going through this too but I hope my experience can help you feel as if you're informed in making your own treatment systems."
>
> -1029, in illness narrative
>
> "What my thought process was was to actually state it in a way in which I wish I would have been told from someone experienced in this! I actually had a friend who went through something which absolutely caused him the same diagnosis! I remember trying to help him out and explaining to him why it's so very important to seek help now rather than wait until his life is falling apart!!!"
>
> -1816, in freeform feedback

Nearly all of IlNeS's Emotional narratives address the audience directly. Many examples were omitted even from this work's appendix due to their volume. Participants in this partition seemed most invested in the imagined scenario of the prompt; multiple participants made offers to help their audience 'later' or role-played providing contact information. The simulated interactions were highly personal, personable, and encouraging. Participants often noted an attempt to be realistic about their condition while still providing hope to their audience. Many reflected that they had projected their own emotions at similar points in their journey onto the simulation.

6.1.4 Intention is Nuanced

The aim of the proxy audience prompts was to elicit distinct modalities of intent. In practice, no typology may be able to full isolate narrative intentions from one another; the category may always be nuanced. Examples of cases where participants stated an intention outside the audience prompt's corresponding NarraType include:

> "Mental illness is not taken as seriously as physical illnesses but they are equally worthy of help and care."
>
> -3517, assigned the Factual NarraType
>
> "i wanted to capture the heavy emotional weight that came along with the experience: the fear, the physical pain and the isolation"
>
> -6874, assigned the Persuasive NarraType

[70] Psychological explanations for altruism are complex, contentious, and outside the scope of this dissertation.

These cases are noteworthy, but they reflect exceptions to the overall success of the proxy prompt design. (A full list of intention mismatches in the IlNeS data can be found in the Appendix 10.3, in Table 51.)

Though these samples demonstrate overall success of proxy prompts, explicit expressions of intent do not necessarily match the behavior of a text. This is a natural limitation of the survey format; human writers (especially lay writers) do not always fully understand why they write in a specific way. For example, Participant 1561 (assigned the Persuasive prompt) wrote in feedback, "Another thing I wanted to do was describe it as much as possible" while their narrative was characterized by a lack of detail. This is a salient example of the discrepancies that can arise between *stated* and *observed* intention within a text. Despite this, the evidence presented for testimonial/feedback pairs across the IlNeS corpus appear generally reliable as markers of authorial intent.

6.2 Clinician Communications

Patient-clinician communication is an important area of patient care. Research asks how clinicians can better build trust with patients, with a goal to ensure that patients fully utilize healthcare resources, provide clinicians with relevant details for their care, and follow clinical instructions for their best health. A common perception of patient-clinician communication is that problems come from the patient side: patients may be accused of withholding health information, deliberate non-adherence, and more.

However, different studies have supported the idea that patient reticence stems from a variety of factors, including a warranted distrust in the medical system (ex. Donovan et al., 1992). Illness narrative theory explores the idea that illness is a multifaceted experience, and that narrative testimonials about this experience can surface these feelings of uncertainty, vulnerability, and need. *Narrative Illness* provides evidence for this from the author's own clinical practice, where allowing her patients to share their emotional stories helped her to align patients' treatment with their healthcare goals (Charon, 2008). The IlNeS work builds on that perspective, and examines participants' statements about a theoretical clinical conversation: their goals, their priorities, and their preexisting feelings about the possibility of a new patient provider.

The IlNeS corpus collects a range of responses. Some are forthright when speaking to an imaginary clinician, while others show skepticism regarding the medical establishment (see Tables 49 and 50 in Appendix 10.3 for details).

6.2.1 Prompt Validation

As with the survey's other NarraType prompts, the work seeks to validate that participants engaged with the prompt as provided. Because this survey was run in a simulated environment, participant behavior may not exactly match their natural behavior in a clinical setting. This section will demonstrate that the results found appear to approximate true clinical interactions.

> "Hi, and thank you for taking time to meet with me. I would like to give a general overview of my medical history as requested since this is our first appointment."
>
> -4399, in illness narrative

Participants often make direct references to the imaginary general physician, especially through framing statements for their overall narrative. The tone is closest to that of Persuasive responses (respectful and formal), though at times participants will use personable, "emotional" language. A full list of relevant framing statements can be found in Appendix 10.3 under Table 49.

> "I simply put myself in the situation of having to describe my history with chronic pain to a new physician who I had not seen previously."
>
> -1891, in freeform feedback

The feedback response section made frequent reference to active engagement with the imagined scenario of meeting a new general practitioner. A full table can be found in the Appendix, under Table 50.

Clinician interactions were also referenced by two participants who did not receive the Clinician NarraType prompt. These answers (shown in Table 17) provide additional insight into how individuals dealing with illness feel about clinical interactions.

ID	Prompt	Feedback
5142	Emotional	"I also focused on sharing things that actually helped me, like [...] finding a healthcare team I trust"
7599	Emotional	"Finding doctors who actually listen and explain things like you're a human? Non-negotiable."

Table 17: Responses from participants who were not assigned the Clinician prompt, talking about their clinical experiences.

In some cases, participants gave clear and direct insight on how they intended to present the testimonial or timeline. 8084 directly referenced narrative in their freeform response ("I thought about my experiences when I was sick and I wrote it out like a narrative story."), while 8599 expressed intent to organize events chronologically ("I focused on presenting my problems in chronological sequence to resemble a standard medical narrative"). The use of the word 'standard' suggests that to some of this dissertation's participant demographic, chronological ordering may be seen as a common feature of clinical communications.

6.3 Illness Narrative as Personal Expression

> "Those living in chaos are least able to tell a story, because they lack any sense of a viable future. Life is reduced to a series of present-tense assaults. If a narrative involves temporal progression, chaos is anti-narrative."
>
> -*The Wounded Storyteller* (Frank, 2013)

Communicative text exists as a product of both internal and external motivations on the part of its author. Charon, in her work, highlights external motivators (i.e. the need to build a bridge with healthcare providers) while Frank emphasizes internal motivators and limitations: "The stories that ill people tell come out of their bodies [...] [which are] simultaneously cause, topic, and instrument of whatever new stories are told" (Frank, 2013). Hard limits exist on human recollection (see Section 3.3.2) and many illnesses feature memory loss and diminished cognitive capacity as symptoms. The trauma of illness can likewise reduce memory capacity. Illness narrative is therefore a genre especially prone to perturbations in event chronology, and as Piper et al. (2022) defines 'narrative' as dependent on a sense of *sequentiality*, such perturbations are especially stark within this domain.

Yet, the attributes of *experientiality* and *world-making* are also factors of narrative. Past chapters have explored the impact of external motivators on sequentiality in the IlNeS corpus; this section moves to the other two elements of Piper's narrative framework as motivated by the internal needs of individual participants. Illness narrative, as a genre, derives from "the

need of ill people to tell their stories, in order to construct new maps and new perceptions of their relationships to the world" (Frank, 2013). Storytelling, in the field of illness narrative, is an active tool for processing illness–a canvas for reconstructing the experience and world that forms during these traumas. Using the rich data provided by IlNeS's raw testimonials and freeform feedback, this section aligns observations from the corpus with past frameworks of illness narrative. It seeks to answer a question about the illness narrative sub-genre: **What does the IlNeS data suggest about the internal motivators behind illness narrative?**

Data collected by IlNeS supports or further elaborates on the following existing theories about illness narrative:

1. Illness narrative serves a *unique social function* to people dealing with chronic or acute health conditions.
2. Telling stories about illness experiences has a *positive impact* on the speaker
3. Telling stories about illness helps speakers to *mentally reconstruct* a chronology.
4. A prevalent element of illness narrative is the speaker framing themself as a *moral actor*

Overviews are discussed below, and full tables of statements supporting each theory can be found in the appendix under Section 10.3.3.

6.3.1 Unique Social Experience

> "I was prepared to tell an emotional story about the journey I've been on for the last 10 years."
>
> -Participant 8811

The Wounded Storyteller notes, "The obvious social aspect of stories is that they are told to someone, whether that other person is immediately present or not" Frank (2013). In IlNeS, many participants referenced the act of illness storytelling as a social experience. Some participants reported that they had been able to tell stories about their experience to others in their life already, while for others this was a new experience. Those who had never told their stories before reported more difficulty in speaking about their experience in the survey–but they also appeared to derive a specific benefit from doing so in the survey setting. This feedback suggests that sharing illness narratives serves a *unique* role in helping individuals to process the experience of illness that cannot be replaced by other types of social interactions.

The response of one participant (8811, quoted above) suggested that this format of illness narrative may be associated with emotional storytelling by default[71]. The participant expressed disappointment with being assigned a non-Emotional NarraType prompt, although they had no knowledge that an Emotional prompt existed. It may be that opportunities to write *emotional* testimonies about illness are particularly desired over others–to further quote 8811: "You [sic] prompt about what to write kinda threw me off."

This trend within IlNeS supports the idea that introducing patients to opportunities to share their illness narrative specifically improves mental well-being, and fills a need that is not met by other forms of emotional outreach.

[71] Supported by other works–Rini et al. (2014) found that emotional writing exercises had positive impact on mental health among recovering cancer survivors.

6.3.2 Sharing Stories Leaves Impact

As a field, illness narrative's central thesis is that telling stories about one's own experience allows the teller to heal from trauma. Proponents (like Frank and Charon) assert that it can allow a storyteller to regain agency over a turbulent period in their life; IlNeS includes some testimonial evidence to support this. Unprompted, multiple participants within the study express feelings of relief, pride, or self-acceptance related to the retelling of their story.

However, other participants expressed mixed ad even negative responses to the act of sharing about illness narrative. Some examples of each are found below. The full collection of relevant excerpts can be found in Tables 53, 54, and 55 within Appendix Section 10.3.3.

> "I had a proud feeling in me. I felt proud that I overcame such a stressful time in my life, and am now discussing it with a complete stranger." -9101
>
> "It was kind of tough reliving it, but kind of freeing at the same time." -3441
>
> "I was feeling rather uneasy while writting [sic], I find it difficult when it comes to talking about some of the dark moments in my life." -8760

There is clear variety in responses. When considering storytelling as a therapeutic tool (similar to trauma-based psychological 'debriefing'), a participant's initial frame of mind is important. The field of illness narrative generally positions storytelling as a positive coping mechanism– despite this, studies have held that this use of narrative does not benefit all patients. For some, storytelling may hinder recovery. A systemic review of debriefing reported an "overall neutral" impact from debriefing practices, and hypothesized that "the intense reexposure involved in the PD can retraumatize some individuals without allowing adequate time for habituation, resulting in a more negative outcome" (Rose et al., 1998).

IlNeS results support this caution: the data includes cases where the re-telling of illness narrative produced some distress communicated to us by participants. (See 5.2.1 for resources given to them in the event of serious emotional harm.) Most survey participants who reported negative feelings completed the full survey; one participant (8612) who shared experiences with post-traumatic stress disorder did not. It is possible that those with especially traumatic illness conditions are more likely to have a negative response to storytelling-style debriefs. Note again that resources for emotional support were provided to participants alongside the consent form and survey. Multiple warnings about the potential for emotional distress were also shown to participants through flyers and Prolific before they could enroll. Participants who chose to take the survey despite these warnings were engaging in self-selection, based on a belief that the work would *not* cause them personal harm. Therefore, it is likely that the larger patient population has a higher proportion of individuals who would exhibit distress from revisiting their experience through story than were observed in IlNeS.

For the positive impact cases, the element of narrative most closely associated with happiness and pride was the patient's *ability* to tell their own story. This ability was framed as a sign of having fully overcome their illness. In the traditional illness narrative framework, storytelling is the action which causes emotional healing. IlNeS analysis may suggest the opposite: that individuals come to terms with their condition first internally, and the stories they tell simply reflect that inward growth. This could explain why distress is such a common feature of illness narrative despite the practice being presented as inherently healing; whether an individual is helped or hurt by this type of debriefing depends on the stage of emotional recovery they currently inhabit.

6.3.3 Recollection

> "I was thinking about the look on my dad's face as I turned blue and that he had a cigarette in his hand. [...] I can almost smell his cigarettes. Memories are freaking weird." -8196

Illness narrative is hypothesized to be motivated by a need "to construct new maps and new perceptions of [the narrator's] relationships to the world" Frank (2013). This perspective is shared by cognitive theory and neuroscience, which agree that recollection includes a step of 'revisiting' important to the conceptualization of events having a chronological order (Buckner et al., 2001; Byrne, 2010). This process may be facilitated by narrative prompts; in IlNeS feedback, there is a trend of participants taking an active step to revisit events, or to experience the process of recollection unconsciously. This process is often *vivid*, with certain participants experiencing associated physical feelings and sensory details.

In the broader domain of narrative, the vivid visualization that occurs in recollection may relate to the *world-making* identified by Piper et al. (2022). In recalling sensory detail, the setting of the illness experience is being reconstructed. Frank asserted this process is common to storytellers, and that this recollection can produce distance between storyteller and their experience which reduces the confusion which had been experienced in the moment. He described the inherent tension of illness narrative as such: to "reduce the body to being the mere topic of the story" is both alienating but can provide some manner of relief (Frank, 2013). One IlNeS participant (6535) speaks to these conflicting feelings, noting "I really have to kind of view [the experience] as a series of events or a story just to keep myself from getting anxious and disgusted with myself." The participant is both obliged (they *"have to"*) express their experience in a certain way, but the storytelling format also *allows* them to move past anxiety and shame.

Another (who produced a timeline which fully matched chronological ordering) called out 'sense-making' explicitly as a motivation for their narrative:

> "I just really wanted to try to give a cohesive picture and timeline of how the whole process went, because sometimes it was a real blur [...] I really have to kind of view it as a series of events or a story just to keep myself from getting anxious and disgusted with myself." -6536

This highlights another distinct feature of time and the experience of illness: the stressful nature of illness impedes clear recall. 6536 outlined an intent to provide a chronological narrative[72] and directly contrasted it against feelings of confusion. They used the freeform feedback section to step again through their own internal timeline ("I first saw my neurologist in [MONTH], saw the opth in [MONTH], and now it's [MONTH]."), suggesting that construction and *re*-construction of the underlying content of their illness experience helps in some way with comprehension.

The IlNeS corpus also identified cases where participants felt that writing testimonials did not help with sense-making. For example, 5345 noted that they "[tried] to remember what happened back then," but had difficulties due to cognitive impairments from illness: "I still have a little brain fog." Such cognitive impairments are expected in the chronically ill population, as many factors of stress, trauma, and illness can impact memory (Van der Kolk, 1998). *The Wounded Storyteller* suggests that illness narratives may feature disorder for deeper emotional reasons. Section 5.1.2 of this dissertation discusses the **chaos narrative**, a subtype of Frank's illness narrative framework where traditional understandings of time in narrative break down: the

[72] Despite being given a Persuasive prompt. It is discussed in Section 5.4 that true chronological ordering is uncommon among that NarraType.

trauma of the situation actively prevents sense-making. "I struggle with remembering what happened during that time," said 2451. Despite stated confusion, 5345 and 2451 demonstrated highly *ordered* narrative timelines. This would seem to contradict Frank's narrative taxonomy; chaos narratives are meant to reflect an *internal property* of the narrator, and are presented as overwhelming the usual structure of narrative.

A potential explanation for this discrepancy is the NarraType prompts. In the cases above, both participants had been assigned a Factual prompt. There is an overall association between Factual narratives and chronological timelines (discussed further in Section 5.4); it could be that participants given this prompt exert greater effort to overcome the chaos of their internal narratives. The act of event annotation may help participants to further make sense of their own narrative–if one cannot tell the entire story in a chronological order, they will focus on the details they are clearest on. In this way, a chronological timeline can be assembled after the fact from a narrative that would otherwise be *chaotic*. As Frank asserts that narrative storytelling is often done to regain feelings of control over tumultuous memories, these examples may actually reinforce his central thesis: even when individual participants admit confusion over their own timelines, they use the tools of storytelling to create sense where it was not previously.

The full set of responses relevant to recollection can be found in Table 56 in Appendix 10.3.

6.3.4 Moral Element

> "I had dome [sic] everything that I was supposed to do, in the order prescribed, but one day, none of that mattered." -1334

In his examination of illness narrative, Mike Bury presents a sub-type of illness narrative known as the **moral narrative** (Bury, 2001). He proposes a moral justification process that often occurs when discussing or disclosing illness to others. In many societies (ex. the United States, where IlNeS collection takes place), illness or disability is seen as a personal failure. In some cases, their social sphere may hold beliefs that illness is a judgment from religious powers, a natural consequence of one's lifestyle, or the result of treatment non-adherence and otherwise stigmatized 'unhealthy' behaviors. A similar societal belief is that the chronically ill or disabled exhibit 'self-pitying' behavior; the audience may believe that the speaker's illness is not their own fault, but still interpret the speaker's attitude as disproportionate to their situation. Discussion of their condition and the realities of it are thus interpreted as dramatic and attention-seeking, out of step with reality.

Bury's definition of **moral narrative** as a common and prevalent subtype of illness narrative assumes that this attitude towards ill individuals is widespread, and therefore that speakers discussing their experiences with illness react by positioning themselves as a moral actor–someone who did not deserve their illness, someone who reacts appropriately to their hardships, and someone who can thus be sympathized with. While the narrative attribute of *experientiality* is common among self-testimonials (as the author by default includes themselves as an agent acting within the narrative), the moral narrative represents a case where the experiential elements of the text are direct and deliberate.

In IlNeS development, the taxonomic classification of a moral narrative was noted to be narrow and highly specialized to the medical domain. To produce results that were more generalizable, the NarraType typology included all narratives where speakers are attempting to influence the audience towards a goal as **Persuasive**. It is nonetheless interesting to explore elements of Bury's moral narrative within the IlNeS corpus. Three trends within the survey data seem to align with Bury's approach:

1. An explicit statement within the text that the speaker's illness is not their fault.
2. An avoidance of emotional language to present oneself as not "attention-seeking"
3. An admission within the text that the speaker's illness *was* their own fault.

Trend 3 may seem at odds with the intention of a moral narrative (to present oneself as a moral actor worthy of sympathy). However, an admission of fault may also serve this same goal, functioning as a sort of *negative moral narrative*. If an individual believes they are responsible for their own illness, framing themselves as "taking accountability" could serve to elevate the speaker's moral position above others with the condition who do not.

(**An important note**: this dissertation is not attempting to attribute blame to any of our participants for their illness. The IlNeS design gave participants the freedom to choose the content of their narrative and the freeform feedback they provided. This analysis simply chronicles cases where participants volunteer their own beliefs about the cause of their condition.)

Examples of each type of moral statement are shown below. The full collection of excerpts may be found in Table 6.3.4.

> "I had dome [sic] everything that I was supposed to do, in the order prescribed, but one day, none of that mattered." -1334
>
> "I was aiming to be honest, realistic, and reflective, and not dramatic or excessively emotional." -0243
>
> "The quality of life is very poor and I realize I have no one to blame other than myself." -8812

Some additional references to morality did not fit the trends above:

> "The toughest part of the diagnosis is the social stigma. Do not be asahamed [sic]." -1236
>
> "Economic & ethical appeal: I combined moral language (dignity, potential) with an investment rationale ('every dollar invested...'), because policymakers respond to both." -9460, who directly linked the moral argument to their intent to persuade

In Chapter 8, the dissertation runs sentiment analysis using LIWC dictionaries to identify if there is a quantitative association between moral language and any of our NarraTypes. Further discussion can be found there.

The Collective Nature of Illness:

It is worth noting that, for all moral trends, participants rarely framed themselves as any *more* or *less* moral than others with this condition. In the positive moral narrative examples, participants either do not mention others with the condition or attribute the same moral character to everyone in their group ("no one asked to be born sick"). In the negative moral narratives, the actions of all in the group are given the same moral weight ("it is due to our decisions").

6.4 Discussion

One research question of this dissertation asks: **How is the behavior of short-form illness narratives driven by the factor of authorial intent?** IlNeS breaks this down into the following discussion questions: DQ1) What can the testimonials in IlNeS tell us about existing

theories in the field of illness narrative? DQ2) What do these testimonials tell us about how patients communicate with their clinicians?

DQ1. Illness Narrative

IlNeS data contains trends which confirm certain theories posited by the field of illness narrative: that illness narrative is perceived as a unique type of communication by participants; that the telling of illness narrative can be a powerful, healing experience; that moral statements are prevalent in illness narrative.

There is also evidence present to refute certain theories. Though illness narrative *could* be a healing experience for participants, it was not universally so. Some participants reported mixed or even negative feelings about re-telling their narrative. Based on these results and the freeform feedback provided by participants, it may be that illness narrative is not a catalyst for healing but a reflection of it; individuals who are more at peace with their illness journey would therefore feel happier after giving testimonials. Note that our recruitment and consenting process included warnings that recalling illness experiences could be traumatizing. It is expected that some potential participants who might have extremely negative reactions to providing an illness narrative self-selected out of the study–what is seen in IlNeS may skew more towards those with positive feelings than is true for the general population.

DQ2. Clinical Text

Many IlNeS statements discuss or reference clinical communication. Some participants convey desire to fully inform their imagined clinician about their condition, suggesting a cooperative attitude and trust in their provider. Others note skepticism, referencing strategies they consider necessary to have a productive interaction with a new provider. In some cases, outright frustration was noted, and these statements were not wholly isolated to the Clinician NarraType. Even when medical providers are not referenced specifically, participants were often thinking about their clinicians when they thought of illness.

6.5 Conclusion

The IlNeS data collection contributes a substantial corpus of patient testimonials with rich data for the field of illness narrative. Qualitative analysis of these works allows theories about illness narrative to be evaluated against multiple data points. The freeform feedback response form is especially useful in analysis, as it provides a direct link between the behavior of a testimonial text and the motivation behind this behavior. This data validates the NarraType labels used to partition IlNeS and contributes deep insights into the experience of illness, seeking support, and clinical care.

7 MulCo Model

In this chapter, the dissertation outlines an end-to-end TEO/ETRE prediction pipeline. This work utilizes the Distilling **Mul**ti-Scale Knowledge via **Co**ntrastive Learning model (called **MulCo**), a project developed in collaboration with Yao et al. (2024) for broader TEO learning.

MulCo differentiates itself from prior TEO/ETRE models through its use of *knowledge co-distillation* to surface and preserve useful features across multiple layers of text. The work notably does not use features that were not present in prior models; rather, the co-distillation method of knowledge synthesis enables the model to better leverage existing data. This approach was motivated by study on TIMERS, an at-the-time SOTA model within the field which combined local context and structural information.

The success of MulCo further motivates the approach in this dissertation's novel TEO/ETRE modeling experiments: where MulCo demonstrates that knowledge co-distillation can effectively synthesize insights from useful features across multiple layers of text, IlNeS's NarrType data introduces a new layer of knowledge for distillation.

I was a co-author on the original paper; my role was to contribute deep understanding of the task to motivate the technical work. I designed experiments to demonstrate and justify the technical decisions behind the model architecture, and provided qualitative analysis of the mechanisms through with MulCo improved upon SOTA. This work motivated the exploration of illness narrative within this dissertation, and the construction of the follow-up model variant MulCo-Int.

This chapter will present the following:

- A technical understanding of MulCo and a variant produced for Wu et al., 2025 which expands the model to other "cross-model problems" (i.e. tasks which rely on multiple layers of a text, or multimedia texts).
- Qualitative analysis from the original MulCo paper and its implications for expansion of the model to the IlNeS dataset.
- Qualitative insights from IlNeS analysis. This information will be placed within the context of the MulCo model to explain how authorial intent can be incorporated into the architecture to produce new model variations for the **TempR-MInt** project.
- Results of technical experiments and analysis.

7.1 Motivation

The focus of the dissertation implements a type of narrative analysis which is rarely applied to statistical modeling[73]. To bridge that gap and bring the insights from qualitative narrative work to TEO/ETRE requires a model with a specific type of architecture. MulCo represents a competitive, state-of-the-art knowledge-distillation model. It was developed for the task of TEO/ETRE, which benefits from distilling information across multiple textual layers (see Section 2.1.2 and its discussion of short- versus long-distance event-pair cues), but also has proven to show utility for other "cross-layer" tasks (Wu et al., 2025). A question driving the dissertation research is whether the narrative element of text constitutes a separate "layer" of language, such that it fills gaps in extant NLP tasks like TEO/ETRE. The work considers MulCo an ideal model with which to explore this question.

[73] See Atkinson (2009).

The dissertation examines two implementations of MulCo: the original, designed for TEO/ETRE in the traditional journalism domain using standard benchmarks (TBDense, TDDiscourse-Auto, TDDiscourse-Manual, and MATRES), and TempR-MInt, a set of new implementations of MulCo directly scoped for the IlNeS corpus and the integration of authorial intent. This section outlines the original iteration of MulCo as it was first implemented.

7.1.1 Task Definition

This section returns to the TEO/ETRE task definition given in Chapter 2 and reiterates specific details important to MulCo model development. Prior modeling successes in three categories (encoder models, structural models, and other) are detailed in Chapter 3.

TEO/ETRE takes as input two **event mentions** from within a text and predicts their order within real time. The four datasets in use within this task are TimeBankDense, TDDiscourse-Auto, TDDiscourse-Man, and the MATRES corpus (Ning et al., 2018). Note these corpora contain content only within the domain of *journalism*, which limits models' utility for general prediction[74].

TEO/ETRE defines five meaningful labels for the temporal relationship between two events. These labels can be said to have two sub-properties of *order* and *overlap*:

- **Before:** If $Rel(E_A, E_B) = before$, the spans of time for which the two events take place do not overlap, and E_A occurs before E_B. The order of mentions in the event-pair is key; if $Rel(E_A, E_B) = before$, $Rel(E_B, E_A) = after$.
- **After:** If $Rel(E_A, E_B) = after$, the spans of time for each event do not overlap, and E_B occurs before E_A.
- **Simultaneous:** If $Rel(E_A, E_B) = simultaneous$, the spans of time for each event fully overlap (e.g. both start at the same time and end at the same time). This label is symmetric; if $Rel(E_A, E_B) = simultaneous$, so is $Rel(E_B, E_A)$.
- **Includes:** If $Rel(E_A, E_B) = includes$ Event B, there exists some incomplete overlap between the spans of time when each event occurs, and A begins before B does. If $Rel(E_A, E_B) = includes$, the label for $Rel(E_B, E_A) = isincluded$.
- **Included By:** If $Rel(E_A, E_B) = isincluded$, there exists some incomplete overlap between the spans of time for each event, and E_B begins before E_A does.

The above five are "meaningful" labels for the TEO task. However, datasets TimeBankDense and MATRES include the additional label:

- **Vague:** This annotation label is assigned to event-pairs for which the original annotator could not make a determination from the five meaningful labels.

This label is used in TimeBankDense by Cassidy et al. (2014) and MATRES by Ning et al. (2018), two datasets traditionally considered baselines for TEO/ETRE. In modeling, the *vague* label marks all event-pairs considered "difficult" for a human annotator working with standard schemas (see Section 4.2 for more on this label's impact). All other labels, in addition to the semantic meaning they encode, implicitly mark the pairs in question as "easy" for human annotators to determine ordering. This creates a two-tiered hierarchy of pair labels, which in theory could help train models to surface the pairs for which it can make the most confident

[74] A motivation behind the IlNeS–by expanding TEO/ETRE to illness narratives, new work can identify useful features which may be absent from news article text.

predictions. In practice, models trained on TimeBankDense and MATRES have a strong prior class bias towards the *vague* label, especially when faced when challenging event-pairs. This is problematic for any use-case of actual temporal processing where a model is expected to make real predictions about even difficult event-pairs.[75]

The TimeBankDense corpus uses the span of text between two mentions (called in this chapter of the dissertation 'context') to differentiate **short-distance** from **long-distance** event-pairs, and included only short-distance pairs. Events within the same sentence or neighboring sentence are considered short-distance, events further apart are long-distance. As a reminder of this important distinction, examples of each may be seen in Figure 58.

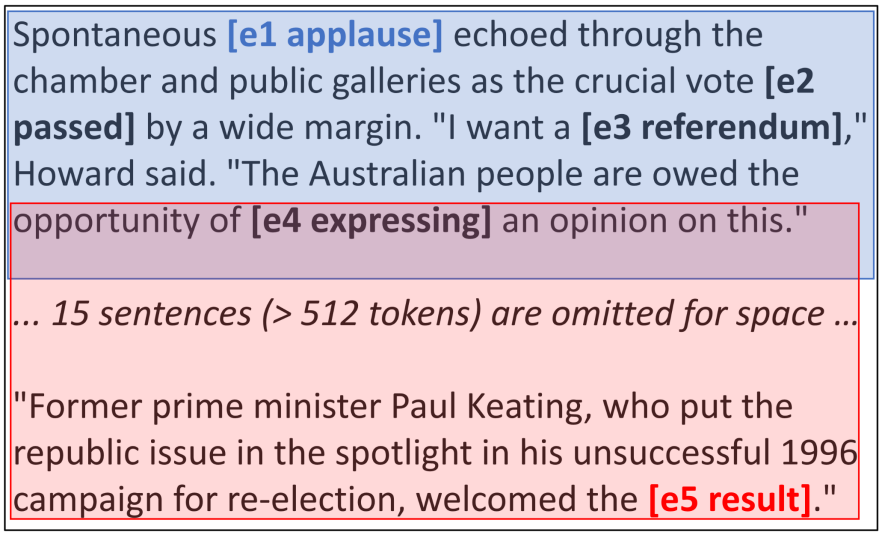

Figure 58: Short-distance pairs have a small content window (blue) and long-distance context (red) is much larger.

The distinction between short- and long-distance event-pairs has implications for TEO/ETRE model architecture. If the example of Figure 58, "(e1) applauded" and "(e2) passed" represent **short-distance** pairs. A human reader can intuit the temporal relationship between the two events solely by looking at the interconnecting context–implicitly, so can a machine learning model using a context encoder (for example, BERT). TDDiscourse-Auto and TDDiscourse-Manual, introduced later (Naik et al., 2019), add long-distance pairs to the TEO/ETRE task which changed the fundamental nature of this reasoning task. Consider the following **long-distance** pair: "(e1) applauded" and "(e5) result". There are over 15 sentences of intervening context, only some of which may be relevant to TEO/ETRE. A model dependent on bi-directional transformers is likely to fail for these type of event-pairs, due the larger context window; additional architecture is required to compensate.

7.1.2 MulCo Motivation

The design of MulCo derived from a motivation to build deep understanding of the needs of short- and long-distance event-pair prediction. Current TEO/ETRE finds long-distance pairs more challenging than short-distance, likely due to an early focus on TimeBankDense and

[75] Note additionally that the MATRES dataset also omits the labels *includes* and *isincluded*.

MATRES (both short-distance datasets). MulCo launched a set of experiments on at-the-time SOTA models, extracting performance data across short- and long-distance pairs. The evaluation performed in MulCo sought not simply to identify *where* current models failed, but *why*. Through this, the work believed that existing architectures could be more effectively leveraged to truly capture temporal reasoning.

Human knowledge is essential for TEO/ETRE; this task relies on multiple layers of textual knowledge which are not simple to pass along to a model. Chapter 2 discussed distinct heuristic signals which can communicate to human readers temporal information. That time can be encoded on many distinct streams of a given text explains why temporal reasoning and TEO/ETRE are such challenges for models–when building ML models, the *layer of text* where the information is located informs effective model architecture. Consider:

- **Short-distance context:** Ex. "Anne headed to her car, after grabbing an umbrella."

  Here, the ordering of the events "headed" and "grabbed" is explicitly cued by the adverb connecting the two clauses. This is a short-distance contextual time cue, and plays a significant role in how most TEO/ETRE models approach the task.

- **Semantic time cues:** Ex. "The merger was finalized on January 15th."

  By using absolute and relative references to the date-time of a given event, an event can be placed in context with other time cues in a text. This method of encoding is a global phenomenon and can be applied to event-pairs no matter their distance in the text.

- **Cultural knowledge:** Ex. "John iced the cake with frosting he made while it was in the oven."

  Common cultural 'scripts'[76] may contain information about temporal ordering; here, the audience likely knows that a cake must go in the oven before it can be iced.

- **Logical entailment:** Discussed in Chapter 2. The logical entailments of text represent a specific type of external knowledge, not encoded in the text itself.

- **Genre-based structural constraints:** Certain domains enforce specific structural norms. For example, clinical appointment records often have a section discussing patient medical history, which captures events at the start of a patient's personal timeline.

These types of cues can be further sorted into local (or *short-distance*) cues and global, structural (*long-distance*) cues. Models specialized towards short-distance cues respond best to event-pairs whose temporal information is located in the local context. However, they underperform on event-pairs that rely on long-distance temporal cues. The success of integrated BERT/GNN models in prior TEO/ETRE work suggest that graphs capture an element of temporal ordering which cannot be encoded by even an advanced bi-directional transformer. Put simply, there are distinct forms of knowledge contributing to temporal ordering, and accounting for these multiple layers of knowledge are necessary to build a comprehensive TEO/ETRE prediction model.

This informs the approach this dissertation takes with MulCo: if the BERT and GNN embeddings each capture distinct types of temporal information, a model's understanding of time can be improved by better distilling that knowledge across the two encoders. In this section, the work details error evaluation and similar experiments run by the MulCo team to identify best next steps for the model.

[76] See Schank et al. (1975).

TIMERS and the MulCo Approach:

MulCo's approach to TEO/ETRE was directly inspired by the TIMERS model produced by Mathur et al. (2021), which directly sought to encode multiple layers of temporal information from within input text. Prior to TIMERS, most SOTA TEO/ETRE models processed textual input in a linear-ordered form–context between pairs was fed to encoders such as BERT (or more efficient long-distance transformers, e.g. Longformer). The work calls this input 'linear' because text is input as a contiguous linear stream. Though attention mechanisms allow important elements throughout the text to be given high weight even across distance of bi-directional transformation ultimately preserves the underlying ordered sequence of text.

TIMERS chose instead to represent structural information through structural models, explicitly encoding hierarchical relations across text with the inclusion of Graph Neural Networks (GNNs). This made TIMERS one of the first multi-stream models in the TEO/ETRE space. Its architecture directly corresponded to certain layers of temporal information referenced earlier in this section. Local context cues (most useful for short-distance event pairs) were captured by the BERT layer, and structural cues (better for long-distance pair prediction) captured by the GNN layers.

The architecture of TIMERS is presented in Figure 59. Its 3 GNNs introduced to the traditional BERT encoding are: a **syntactic** graph which encodes relationships like word order, sentence order, and grammar across the full text; a **semantic** graph encoding exact dates extracted from text; and a **rhetorical** graph tracking steps made between clauses throughout the document.

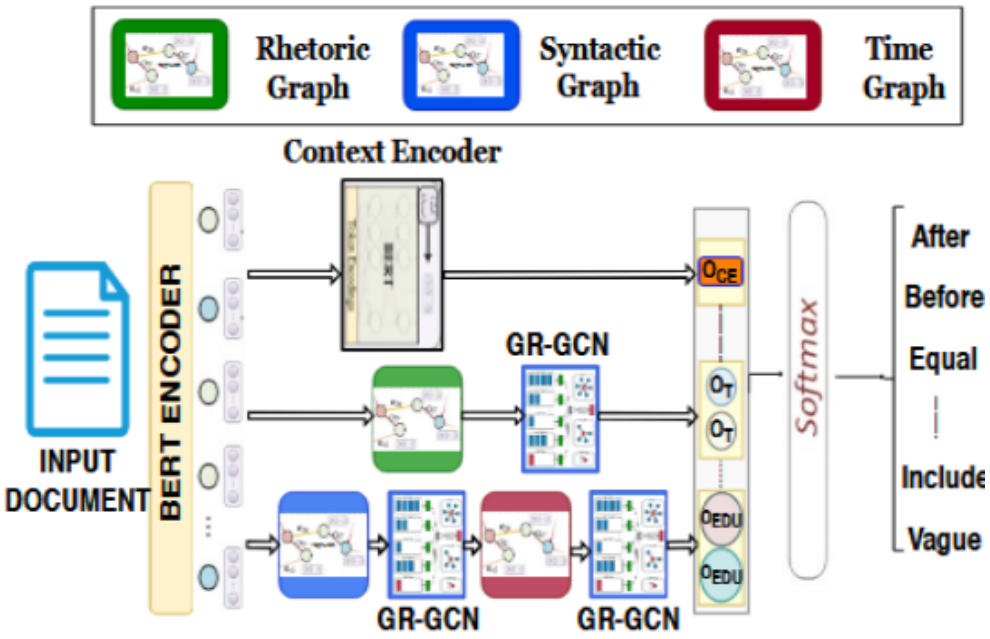

Figure 59: TIMERS pipeline. Image originally printed in Mathur et al. (2021).

Overall, the inclusion of GNNs improved the model above existing SOTA (see Table 18), even for datasets which contained only short-distance pairs.

This work by TIMERS inspired the '**Mul**ti-Scale' approach of MulCo. TIMERS saw modest

	MATRES	TB-Dense	TDDAuto	TDDMan
SP+ILP	76.3	58.4	46.1	23.8
BiLSTM+MAP	75.5	64.5	57.1	41.1
Longformer	-	-	66.8	44.2
TIMERS	82.3	67.8	71.1	45.5

Table 18: F1 scores of prior TIMERS model compared to select context-enabled SOTA.

improvement compared to SOTA of the time by explicitly encoding layers of temporal information from within a text. But whether or not TIMERS represents a true step forward depended on how much overlap exists between its two base models. This raises a key research question: do GNN embeddings *accentuate* the performance of existing BERT pipelines or do they *solve unique problems* with the TEO/ETRE task space?

MulCo's research team formed the following hypothesis: the different models within TIMERS do capture distinct features from with the text (i.e. layers of temporal knowledge). Further, a path forward could be found using TIMERS as a base by improving **knowledge synthesis** between BERT and GNNs. To test this hypothesis, the MulCo project examined the performance of the BERT and GNN implementations from TIMERS in isolation. These experiments used the TEO/ETRE benchmarks TDDiscourse-Auto and TDDiscourse-Manual (both datasets with distinct mixtures of short- and long-distance pairs).

The aim of these experiments was to examine the *specific* pairs from each test set which were correctly predicted by each model in isolation. If the models correctly predicted for *distinct subsets* of pairs in the overall dataset, that demonstrates that the individual models capture fundamentally different layers of temporal knowledge.

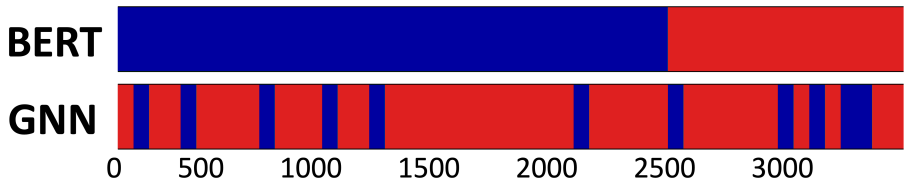

Figure 60: Distribution of predicted pairs that are correct (blue) and incorrect (red) for the TDDiscourse-Auto corpus per model type.

Figure 60 shows the results on TDDiscourse-Auto. This data demonstrates that there is some overlap in the event-pairs which are predicted by BERT in isolation and the GNNs in isolation. However, there are clear subsets of the data where only BERT or only GNN is able to extract enough relevant temporal knowledge to make a prediction. In qualitative examination of these subsets, the MulCo team noted two particular instances where GNN succeeds compared to the BERT model that supported the research hypothesis:

1. Pairs ten or more sentences apart: Event-pairs whose mentions are ten or more sentences from one another are most likely to be successfully captured by GNN and not by BERT. This suggests an upper limit to BERT's ability to encode meaningful context, but that features exist which can nonetheless be captured through the GNN model.
2. Same-sentence pairs separated by some type of reporting verb (e.g. "said", "reports", "announced"): These are short-distance pairs, and intuitively should respond well to

short-distance context encoding. In practice, these pairs are likely to be misclassified by the BERT model–possibly because there is no regular pattern to when events described in a report occurred compared to the time of reporting. GNN models, by contrast, have shown to be able to capture some examples of this pair type, showcasing its ability to extract knowledge beyond immediate linear context.

The results of this experiment support a general trend observed in TEO/ETRE: structural information cannot be captured using only BERT and requires distinct architectures (see the progression of TEO/ETRE types in Chapter 3). Further evidence was found when comparing the performance of BERT models and GNN models across the two TDDiscourse datasets. Table 19 shows this data, and Table 20 splits results for TDDiscourse-Auto into short- and long-distance.

(Note that the information captured in these tables represents *unique* predictions for the model, not the total performance. There are pairs correctly predicted by both models omitted from the count.)

Model	TDD-Auto	TDD-Man
BERT	1065 (4258 total)	91 (1500 total)
GNN	105 (4258 total)	170 (1500 total)

Table 19: Unique predictions made per dataset for each model type.

	Short	**Long**
TDDAuto Number of Test	877	2759
Unique Prediction		
BERT	270	795
GNN	22	83

Table 20: Unique predictions made by each model on TDDiscourse-Auto, split by pair distance.

These two experiments reveal specific insights to the MulCo research team. The BERT context encoder under-performs on the TDDiscourse-Manual dataset (which has mostly long-distance pairs) compared to TDDiscourse-Auto (which has a larger proportion of short-distance pairs in its mixture)[77]. Meanwhile, the GNN architecture alone captures only a small slice of the mixed TDDiscourse-Auto dataset, but almost doubles BERT's unique-pair performance when used to predict for TDDiscourse-Manual. This suggests that BERT will always under-perform datasets with high proportion of long-distance pairs, and that GNNs have some potential in filling this gap. However, within the TDDiscourse-Auto dataset, unique predictions for BERT and GNN are achieved at *similar* rates across short- and long-distance pairs. This seems to contradict the idea that short-distance pairs are best captured using BERT encoding. Further, the performance of TIMERS on benchmarks can be primarily attributed to the BERT encoder–even when only predictions for long-distance pairs are taken into account. GNNs do not perform well on datasets without a significant majority of long-distance pairs. (See Table 21, below.)

The relative under-performance of GNNs in this experiment could be explained in a number of ways. For one, GNNs are relatively recent introductions to the task of TEO/ETRE; where directional transformers have been used for this purpose for decades, only a few models have

[77]The expected value for correct unique TDDiscourse-Manual predictions (assuming the null hypothesis that there are no differences in model behavior across datasets) is 375 unique predictions for the BERT model and 36 for GNN.

Model	F1
TIMERS	71.1
BERT (short pairs only)	36.0
BERT (long pairs only)	70.7
BERT (all pairs)	64.0
GNN (short pairs only)	32.8
GNN (long pairs only)	43.8
GNN (all pairs)	41.7

Table 21: Comparison of BERT alone and GNN alone in re-implementation against TIMERS performance on TDDiscourse-Auto mixed dataset.

utilized GNNs for this task. There has not been much time to fine-tune graph representations of temporal text, and there may be relational representations which better leverage the structural potential of GNNs that have yet to be discovered.

For both model types (BERT and GNN), performance was weaker on short-distance pairs only than long-distance pairs. This may be a problem of definition; follow-up experimentation showed a jump in performance for BERT (short pairs only) when the definition of short-distance is changed to include pairs within a 3-sentence distance. The current definition of short- versus long-distance pairs is ultimately arbitrary. Therefore, it is possible that short-distance linguistic cues may still be useful for some sentences within a "long-distance" range[78]. Note also that these experiments used *training* datasets with both short- and long-distance event-pairs. It was only in the *prediction* step where pairs where split by type. This suggests that there may be short-distance linguistic cues which aid in the understanding of long-distance event-pairs, and structural knowledge which helps predict for short-distance pairs–that is, while the layers of text do capture distinct streams of temporal information, that information does not necessarily map exactly to a simple short- or long-distance cutoff.

The insights captured by these experiments proved useful for MulCo work. It was evident that BERT encoders and GNNs do capture some *unique features* from across text, and there is some observable correlation between these features and event-pair types. While there is some overlap in what each model type can capture, the space of unique predictions suggested potential for a more efficient fusion model than TIMERS. This is where the MulCo project saw potential to improve on the SOTA; it aimed to build an end-to-end pipeline which better leveraged the distinct behaviors of BERT and structural GNNs across training.

MulCo achieved this by replacing the feature concatenation of TIMERS with **knowledge co-distillation**. This allowed the full pipeline to identify the most useful features from each model across all pair types and preserve them across training (unlike simple concatenation, in which useful GNN features were easily overwhelmed by noise from BERT). The results of MulCo compared to prior SOTA (to be discussed in Section 5.4) demonstrated the following, important for validation of this dissertation's fundamental hypotheses:

1. There exists a unique interplay between different layers of temporal knowledge in a text.
2. This knowledge cannot be captured as effectively by a single bi-directional or attentional transformer as by a multi-stream approach.

[78]The MulCo paper (Yao et al., 2024) goes into more detail about the distance cutoff for short- and long-distance pairs, and the reasoning behind retaining the 2-sentence cutoff for modeling during its development.

3. The nature of temporal knowledge in a text across multiple knowledge streams can be surfaced by a well-chosen learning model.

Next, the work discusses a prior contribution of the dissertation (the **STAGE** timex extraction tool) and its potential for integration into MulCo.

7.1.3 STAGE in MulCo

MulCo builds most directly on the architecture of TIMERS (Mathur et al., 2021), a TEO/ETRE prediction models which used syntactic, semantic, and rhetorical graphs to encode structural hierarchies. The **semantic** graph is also known as a **temporal** graph, extracted using timex annotations from the original TimeBank work (Pustejovsky et al., 2003b) and additional TimeML annotations (Pustejovsky et al., 2003a). Time expressions, (or timexes) where previously discussed in this dissertation in Section 3.1.1. They reflect a clear and unambiguous source of temporal encoding, but are often difficult to leverage in NLP.

The TIMERS temporal graph builds nodes for all **events** and for all **timexes** per document. Relation edges link timexes to events, and timexes to each other. Because absolute time expressions (i.e. represents an exact date in real-time) can be heuristically ordered relative to all other absolute time expressions, this method builds a densely-interconnected web across a given input document that is insensitive to the distance between events. From there, graph neural networks can learn about the behavior of event nodes across multiple graph hops. This, in theory, gives the temporal graphs significant power for long-distance pair prediction. However, TIMERS's process of temporal graph construction made a number of assumptions about the state of its input data. The MulCo team believed these assumptions limit the potential of the temporal graph, and proposed a potential improvement using the existing timex extraction tool **STAGE**[79].

When TIMERS linked event nodes to timexes on the temporal graph through a relational edge, it used the criteria that event and timex were located in the same *sentence*. Sentences contain multiple clauses, and it is not certain that a timex in one clause communicates information about events in other clauses. Further, while TIMERS used edges with multiple TEO/ETRE labels to define the ordered relationship between *absolute timexes*, the edges between all events and linked timexes had the value *simultaneous*. In Section 3.1.1, this work noted that time expressions do not always describe the exact moment of a corresponding event; there is grammatical information which orients an event towards a linked timex. The Semantic Temporal Alignment Grammatical Extraction parser, or **STAGE**, is a tool designed specifically to capture this grammatical link and its orientation in time. Different timex parsers[80] do not extract this attribute of temporal orientation, and therefore STAGE presents a unique potential within TEO/ETRE. One simple improvement to the TIMERS architecture could be using STAGE output instead of timex annotations to build the semantic temporal graph used in the pipeline.

(This information could also be extracted through t-links, contained in Pustejovsky et al. (2003b) annotations. However, those annotations are not usable outside TimeBank-derived benchmarks. Any experimentation on new corpora would require new hand annotation or an automated extraction tool, like STAGE.)

[79] Discussed in Section 3.1.1.
[80] Also discussed in Section 3.1.1.

7.2 MulCo TimeBank Implementation

Here the dissertation describes the exact implementation of MulCo used in the original project (Yao et al., 2024). This implementation of MulCo was scoped for the 4 TimeBank-derived TEO/ETRE datasets (TimeBankDense, MATRES, TDDiscourse-Auto, and TDDiscourse-Manual).

Model	**MATRES**	**TBDense**	**TDDAuto**	**TDDMan**
Previous Baselines				
SP + ILP	76.3	58.4	46.1	23.8
BiLSTM	59.5	48.4	51.8	24.3
BiLSTM + MAP	75.5	64.5	57.1	41.1
DeepSSVM	-	63.2	58.8	41.0
Long-Document LM				
Reformer	-	-	65.9	43.7
BigBird	-	-	65.3	43.3
Recent SOTA				
UCGraph	-	59.1	61.2	43.4
TIMERS	82.3	67.8	71.1	45.5
RGST	82.2	68.7	-	-
UF	82.6	68.1	-	-
SCS-EERE	83.4	-	76.7	51.1

Table 22: Previous TEO/ETRE models and their performance on four existing benchmarks.

Table 22 shows the performance of SOTA models at the time of this original MulCo work. The earliest examples utilized short-distance context encoding and performed best on corpora with a bias towards short-distance pairs, with performance dropping for datasets with more long-distance pairs. Long-document language models lacked baseline scores for TimeBankDense and MATRES, but showed improvement compared to short-distance encoders for both TDD-Auto and TDD-Man. Most recent SOTA had begun using combinations of BERT and GNN, which showcased moderate-to-significant improvement across all datasets. For all models so far, performance decreased the more long-distance event-pairs are present in the corpus; this represented the greatest challenge for this task.

7.2.1 Methodology

The MulCo model takes the base of its architecture from TIMERS (Mathur et al., 2021), with minor modifications to the GNN structure. The most significant algorithmic contribution MulCo makes to the task is its **knowledge co-distillation** mechanism, which allows long-distance dependencies to be passed along from the syntactic and semantic graphs to the BERT context encoder. This approach allows the capture of temporal information for event-pair types from all across a document. The full MulCo pipeline is visualized in Figure 61 (below).

BERT Context Encoder:

Like TIMERS and similar prior models, MulCo encodes short-distance context using a pre-trained BERT variant. For each event within the document, a window of $2m$ sentences surrounding the event is extracted and run through BERT. This represents the "context" of the event. For an event-pair, the two contexts of each event are joined together, in the order that these contexts appear in the text–note that for events within m sentences of each other, this concatenation must be modified so that the area where contexts overlap only appears once.

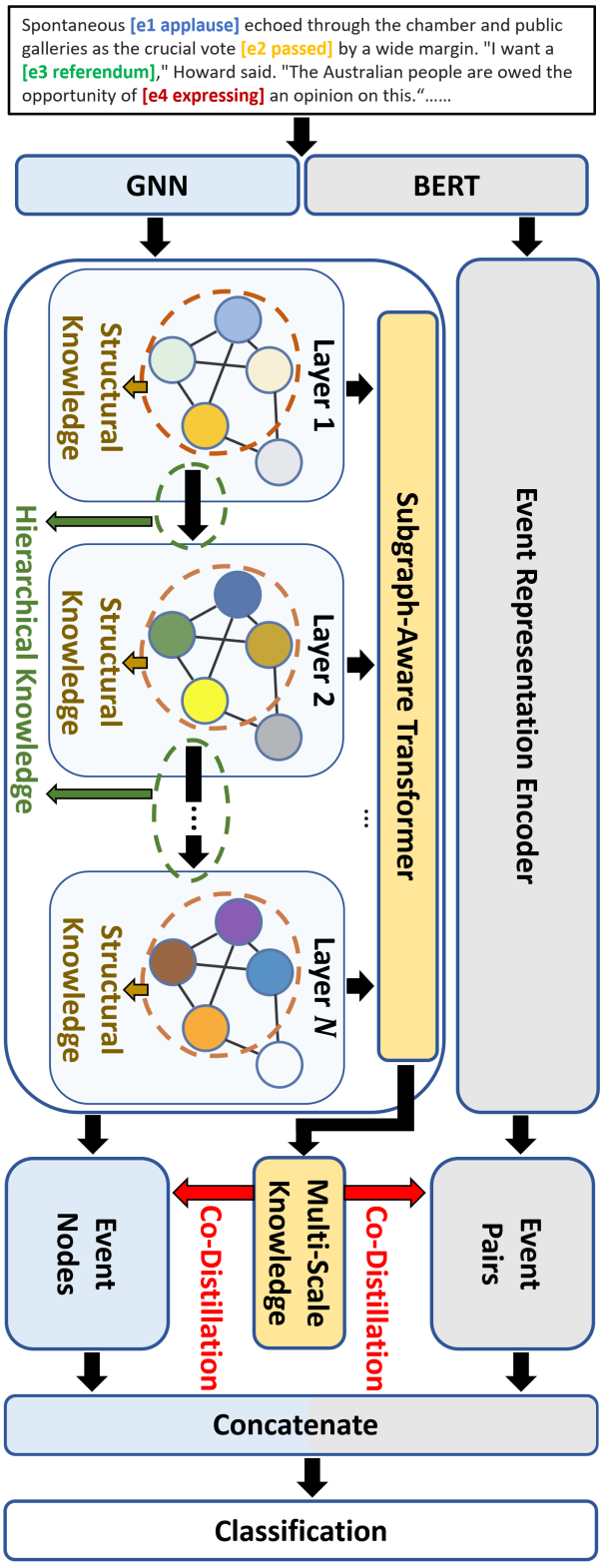

Figure 61: MulCo pipeline.

GNN Dependency Encoder:

MulCo represents long-distance dependencies of the document using two graphs derived from TIMERS (Mathur et al., 2021). Node embeddings are initialized using the BERT context for the extracted event mention.

The syntactic-aware graph creates distinct nodes per word and sentence in the document, along with a top-level document node. Word nodes encode the BERT embedding per word in the document, and sentence nodes average BERT embeddings across the sentence. The edges of the graph encode the following relationships:

1. **Document-sentence:** edges between a document-level node and each individual sentence.
2. **Sentence-sentence:** edges between neighboring sentences, used to preserve sentence ordering
3. **Sentence-word:** edges between each sentence and its constituent words
4. **Word-word:** edges between neighboring words, used to preserve word order.
5. **Word dependency:** uses StanfordNLP parsing to identify grammatical dependencies between words in a sentence.

The nodes of the time-aware graph cover all events within a document, explicit semantic date-time cues (timexes), and a document creation node. Time expressions for this implementation of MulCo are extracted and regularized using SUTime time cue parser (Chang et al., 2012)[81]. The edges encode the following relationships:

1. **Document creation-time expression:** Directed edges which compare given time expressions to the document creation date, encoding before/after and simultaneous relationships. Determinations about the ordering of time expressions are made using date-time mathematical operations.
2. **Time expression-time expression:** Directed edges comparing time expressions to other time expressions with before/after and simultaneous relationships.
3. **Time expression-event:** Un-directed edges linking time expressions to related events. These relationships are determined through grammatical parsing–time expressions that grammatically modify a given event are assumed to describe the temporality of the event.

Knowledge Distillation Across Embeddings:

Section 7.1 discussed experimentation on MulCo's predecessor model TIMERS. These experiments demonstrated that BERT and GNN models in isolation did successfully capture unique event-pairs for TEO/ETRE, but that fusion BERT/GNNs were not necessarily using these distinct elements to their full potentials. SOTA models which concatenated the information provided by both BERT and GNN architecture (such as TIMERS) found some improvement over competing models, but the overall bias of the architecture was towards the information provided by BERT embeddings. This led to salient long-distance dependencies being captured by model GNNs, but discarded in final predictions.

MulCo proposed that to avoid this bottleneck and better retain relevant features across training, a more sophisticated method of fusing the two models was required. MulCo presented knowledge

[81]Though it was determined for the reasons discussed in Section 7.1.3 substituting this parser with STAGE had potential to improve the work further in future work.

distillation as a novel fusion method for this task, with subgraph-aware transformers to preserve structural information and a contrastive co-distillation training approach. This is the primary innovation contributed by MulCo, and it will be described mathematically and conceptually in the following steps:

1. Initial knowledge distillation
2. Structural distillation
3. Hierarchical distillation
4. Subgraph-Aware Transformer
5. Multi-scale distribution
6. Contrastive co-distillation

1) Knowledge Distillation

Knowledge distillation mimics the process of passing knowledge from a human teacher to student. The traditional approach of knowledge distillation features two models: a "teacher" model that is effective but not necessarily efficient for the desired training task, and a simpler "student" model. The goal of knowledge distillation is to train the student model to mimic the teacher's predictions using its less complex architecture–this is typically achieved by minimizing the distances (KL-Divergence) in the embedding spaces of both models. MulCo uses the knowledge distillation objective produced in Hinton et al. (2015) (See Equation 9).

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{N} \text{KL}(\sigma(f_\alpha(\boldsymbol{h_i^{\text{bert}}})), \sigma(f_\beta(\boldsymbol{h_i^{\text{gnn}}}))) \tag{9}$$

Note that $\boldsymbol{h_i^{\text{bert}}}$) and $\boldsymbol{h_i^{\text{gnn}}}$ represent the logits of BERT and GNN event-pair representations and $\sigma(\cdot)$ is a softmax layer. Experimentation found that knowledge flows better from long-distance GNN models to BERT than vice-versa. Therefore, MulCo starts by using this distillation mechanism to pass knowledge from GNN to BERT.

The KL-Divergence objective described above is limited to cases where knowledge directly leads to successful model prediction, and may overlook types of structural knowledge with a more indirect effect on output. To compensate for this, MulCo adds structural knowledge to its distillation function.

2) Structural Distillation

The aim of MulCo is to encode structural knowledge to improve long-distance event-pair prediction. To transfer that structural encoding within a knowledge distillation function, MulCo must encode the topology of multi-hop travel across graph edges (k hops away from a node, or the *k-hop neighborhood*) into the distilled knowledge representation. Denote $\mathcal{G}_i$ as the set of all nodes in the k-hop neighborhood of some node $\sqcap_i$.[82]

$$\ell_{\text{cl}}(t, s) = -log \frac{e^{\text{sim}(t,s)/\tau}}{\sum_q \mathbb{1}_{[q \neq t]} e^{\text{sim}(t,q)/\tau}} \tag{10}$$

[82]Tian et al. (2020) proposed that structural information could be similarly encoded and preserved during knowledge distillation through the use of contrastive learning objectives, explored more in that step.

Similarity function ($sim(t, s)$) is calculated using cosine similarity between t and s, using linear projection to place t and s on the same dimension in cases where they were not already. τ is a temperature parameter. Contrastive loss is minimized for the k-hop neighborhood of event node u_i^1 and u_i^2 using the following equation:

$$\mathcal{L}_{\text{SD}} = \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{G}_i|} \ell_{cl}(\boldsymbol{h_i^{\text{bert}}}, v_j) \tag{11}$$

where $\mathcal{G}_i$ is the set of all node representations in the k-hop neighborhood of event node u_i^1 and u_i^2 such that $\mathcal{G}_i = \mathcal{G}_{u_i^1} \cup \mathcal{G}_{u_i^2}$.

3) Hierarchical Distillation

MulCo encodes hierarchical structural knowledge from the GNN model by stacking L layer GNNS. The hierarchical knowledge distillation follows the loss equation:

$$\mathcal{L}_{\text{HD}} = \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{j=1}^{|\mathcal{G}_i^l|} \ell_{\text{cl}}(\boldsymbol{h_i^{\text{bert}}}, v_j) \tag{12}$$

where $\mathcal{G}_i^l$ denotes the set of l-th layer node representations in the k-hop neighborhood of event nodes u_i^1 and u_i^2 such that $\mathcal{G}_i^l = \mathcal{G}_{u_i^1}^l \cup \mathcal{G}_{u_i^2}^l$.

These equations represent the mechanism of MulCo knowledge distillation. This approach captures many distinct layers of graph traversal (the multi-scale approach), and is therefore inefficient and expensive. To run this knowledge distillation efficiently at multi-scale, additional elements must be added to knowledge distillation.

4) Subgraph-Aware Transformer

To efficiently encode the information across the multi-scale knowledge distillation approach, MulCo utilizes a **Subgraph-Aware Transformer** f_{sat}. Without the pooling effect produced by the SAT, the number of targets in hierarchical distillation would be in the order $\mathcal{O}(N \times L \times K)$ for nodes N, k-hop limit k, and GNN depth L. The SAT functions by computing self-attentive node representations using formula:

$\mathcal{H} \in \mathcal{R}^{d_g}$:

$$\mathcal{H} = f_{\text{sat}}(\mathcal{G}) \tag{13}$$

$$= \text{Pool}(\sigma(\frac{\mathcal{G}W_Q(\mathcal{G}W_K)^\intercal}{\sqrt{d_a}})\mathcal{G}W_V) \tag{14}$$

where W_Q, W_K, and $W_V \in \mathcal{R}^{d_g \times d_a}$ are attention weights, and σ is a softmax function.

5) Multi-Scale Distribution

Using the SAT, MulCo computes a fused structural representation of the set of l-th layer node embeddings in the k-hop neighborhood ($\mathcal{G}_\rangle^\updownarrow$) for event nodes u_i^1 and u_i^2. It then aggregates fused structural representations pulled from the L layer of event nodes u_i^1 and u_i^2 with another use of the SAT to obtain the final layer-wise fused hierarchical representation $\boldsymbol{\mathcal{H}_i}$.

$\mathcal{H}_i$ acts as multi-scale context representation which encodes general, structural, and hierarchical elements of event-pair relations. It is compact, with the number of distillation targets

independent of hops and layers. This allows MulCo to re-formulate distillation equation Eq.12 with the following to achieve efficient **Multi-Scale** Distillation:

$$\boldsymbol{H_i^l} = f_{\text{sat}}(\mathcal{G}_i^l) \tag{15}$$

$$\boldsymbol{\mathcal{H}_i} = f_{\text{sat}}([\boldsymbol{H_i^1}, \cdots, \boldsymbol{H_i^L}]) \tag{16}$$

$$\mathcal{L}_{\text{MD}} = \sum_{i=1}^{N} \ell_{\text{cl}}(\boldsymbol{h_i^{\text{bert}}}, \boldsymbol{\mathcal{H}_i}) \tag{17}$$

Contrastive Co-Distillation:

MulCo differentiates is distillation approach from others with a key element: it does not use a singular "student" model that is learning from a specific "teacher".

The work in MulCo builds from the motivation that both BERT and GNN are learning distinct and unique features about the underlying data. Therefore, the MulCo distillation aligns BERT and GNN embedding spaces to produce a *shared* output layer which, in theory, captures the multiple forms of temporal knowledge of the type detailed in experimental motivations. This use of knowledge distillation appears in prior works (Zhang et al., 2018; Li et al., 2020) as a means of generalizing between two separate but related models, but MulCo combines this approach with multi-scale distillation, which preserves general knowledge, structural knowledge, and hierarchical structural knowledge across both BERT directional embeddings and multi-hop graphical embeddings.

Contrastive representation learning and a stop gradient operator allow the embedding spaces to align without overfitting. The **Contrastive Co-Distillation** loss function is as follows:

$$\mathcal{L}_{\text{CoD}} = \sum_{i=1}^{N} \ell_{\text{cl}}(\boldsymbol{h_i^{\text{bert}}}, \hat{\boldsymbol{\mathcal{H}}}_{\boldsymbol{i}}) + \ell_{\text{cl}}(\boldsymbol{\mathcal{H}_i}, \hat{\boldsymbol{h_i^{\text{bert}}}}) \tag{18}$$

$$\mathcal{L}_{\text{CLF}} = \sum_{i=1}^{N} \ell_{\text{ce}}(f_\phi(\boldsymbol{h_i}), r_i) \tag{19}$$

$$\mathcal{L}_{\text{MulCo}} = \mathcal{L}_{\text{CoD}} + \mathcal{L}_{\text{CLF}} \tag{20}$$

In these equations $\hat{\cdot}$ represents the stop gradient operator (X. Chen et al., 2021) that sets a constant input variable. f_ϕ is a fully-connected layer for label prediction, r_i is the temporal relation for the i-th event pair, and ℓ_{ce} is cross-entropy loss.

This framework using a single loss function $\mathcal{L}_{\text{MulCo}}$ with end-to-end training, unlike prior work in the field (Zhang et al., 2018). This allows **MulCo** output to ensemble separate layers of temporal knowledge across different aspects and scales of the text, simultaneously distilling and optimizing BERT and GNN towards the TEO/ETRE classification task.

The next section will detail the performance of this model compared to baselines, but the potential of the method extends beyond the task of TEO/ETRE. MulCo presents a powerful

Multi-Scale Contrastive Co-Distillation method, which contributes a significant methodology for aligning disparate embedding spaces across machine learning. Wu et al., 2025 would later leverage MulCo architecture to evaluate representational complementarity across a diverse set of multi-modal and text-graph relational reasoning tasks. This dissertation builds on MulCo exactly because of its knowledge co-distillation approach; the dissertation presents the narrative layer of time as a third stream of temporal knowledge in text, and therefore uniquely suited to benefit from multi-scale knowledge co-distillation.

7.2.2 Results

In this section, the work discusses MulCo results on TimeBank corpora.

Corpora:

Of the four corpora which derives from TimeBank (TimeBankDense, Cassidy et al., 2014; MATRES, Ning et al., 2018; TDDiscourse-Manual and TDDiscourse-Auto, Naik et al., 2019), there are three distinct methodologies for scoring the performance of a model. Each will be referred to as **F1** in tables throughout this dissertation, but all models for a given corpus are evaluated using the specific methodology detailed below, unless otherwise specified:

1. TimeBankDense evaluation is performed using six TEO/ETRE labels: the five meaningful relation labels and a *vague* miscellany label. In this evaluation schema, all six labels are treated as positive labels. Models which predict a meaningful label for a pair whose gold-standard annotation is *vague* are considered incorrect, even if the prediction may be correct for the pair's true order.

   This dissertation has discussed the issues with the introduction of a meaningless temporal classifier (see Section 4.2). Following from this work, MulCo chose to perform additional experiments where scores for TimeBankDense are calculated only for cases where $Rel_{gold}(E_A, E_B)$ is a meaningful TEO/ETRE label. This is done to demonstrate how the MulCo model performs under realistic use-case conditions where predictions could be left vague. Comparisons to baselines (as in Table 23, below) include *vague* pairs, while the non-vague analysis is shown later in Table 24.

2. MATRES evaluation includes the five meaningful relation labels and the *vague* miscellaneous category, but *vague* is treated as a negative label rather when calculating F1.

   MulCo performs a second analysis for MATRES performance which excludes the *vague* classifier, found in Table 24.

3. TDDiscourse-Auto and TDDiscourse-Manual only contain five meaningful relation labels in their dataset and do not include *vague* labels at all. As such, all baselines use the same evaluation metric.

F1 Analysis:

Baseline selection for this task covers three types of model:

1. **Context-Encoding Models**: These baselines are relatively old (SOTA as of 2017-2019). At this time, TEO models primarily focused on advancing the implementation of short-distance context encoders. This category includes the following models: SP+ILP (Ning et al., 2017), BiLSTM (Cheng et al., 2017), BiLSTM+MAP (Han et al., 2019b), and DeepSSVM (Han et al., 2019a).

Model	MATRES	TB-Dense	TDDAuto	TDDMan
SP+ILP	76.3	58.4	46.1	23.8
BiLSTM	59.5	48.4	51.8	24.3
BiLSTM+MAP	75.5	64.5	57.1	41.1
DeepSSVM	-	63.2	58.8	41.0
Reformer	-	-	65.9	43.7
BigBird	-	-	65.3	43.3
Longformer	-	-	66.8	44.2
UCGraph	-	59.1	61.2	43.4
TIMERS	82.3	67.8	71.1	45.5
RGST	82.2	**68.7**	-	-
UF	82.6	68.1	-	-
SCS-EERE	83.4	-	76.7	51.1
MulCo	**84.4**	67.5	**77.1**	**55.1**

Table 23: Current SOTA models for TEO/ETRE across all datasets, using evaluation metrics described above. Best results in bold.

2. **Long-Document Language Models**: Though they still use context encoding, this set of models began to explore the use of long-document language models to better capture distant event-pairs. This set includes Longformer (Beltagy et al., 2020), Reformer (Kitaev et al., 2020), and BigBird (Zaheer et al., 2020). Note that for these baselines, only results for TDD-Man and TDD-Auto (the only two datasets featuring long-distance pairs) are available.
3. **Graphical Neural Network Models** Beginning in 2021, SOTA in this field began to integrate GNNs in some capacity. From this category, we include UC-Graph (Liu et al., 2021), TIMERS (Mathur et al., 2021), and RSGT (J. Zhou et al., 2022). Current best-performing approaches overall are SCS-EERE (Man et al., 2022) and Unified-Framework, or UF (Huang et al., 2023).

As discussed, Table 23 uses the *vague* label where datasets include it during F1 calculation. However, this dissertation argues the metric is less useful for models intended to have practical downstream applications than the version that omits it. In real-world modeling, users expect the tool to make meaningful predictions about event-pair order, even for challenging pairs. Table 24 shows the results of training the model only on the five meaningful event-pair ordering labels compared to training the model on these five and the *vague* label. Note that for TDD-Auto and TDD-Man (datasets which exclude the *vague* label entirely) there is no difference in performance, but for MATRES and TimeBankDense there is a dramatic improvement in accuracy (see Table 25 for distribution of all labels per dataset).

Model	MATRES	TB-Dense	TDD-Auto	TDD-Man
MulCo (w/vague)	84.4	67.5	77.1	55.1
MulCo	**90.4**	**85.6**	**77.1**	**55.1**

Table 24: Comparison of MulCo trained and tested on only meaningful TEO labels.

Because SOTA models of the time always included the *vague* label in predictions, these numbers cannot be definitively compared against them. However, it is suggestive that the knowledge distilled during the MulCo process captures real temporal meaning as opposed to just under-

standing distinctions between simple and difficult event-pairs.

	Bef	**Aft**	**Sim**	**Inc**	**IsInc**	**Vague**
TB-Dense	.22	.18	.02	.05	.06	.47
MATRES	.49	.23	.02	0	0	.26
TDD-Auto	.32	.28	.16	.11	.13	0
TDD-Man	.27	.13	.03	.38	.19	0

Table 25: Distribution of labels across datasets. Note that for TimeBankDense, correct prediction of *vague* labels is the largest factor in F1 performance.

BERT vs GNN:

An additional result important to our dissertation is the behavior of the MulCo model beyond simple F1 scores, and how that relates to the individual performance of BERT and the GNN.

Section 7.1.2 showcased the distinct subsets of pairs for which short-distance BERT context embeddings made successful predictions and where only long-distance GNN context did. In the best-case fusion approach, there would be a perfect intersection of correct guesses in the final fused model. As shown in Figure 62 and Table 26, MulCo does not perform as well as a perfect aggregated model, but does capture much of the best-case intersection. It even correctly predicts for pairs which neither original model could predict alone. Though not conclusive, these results do suggest the model is succeeding in its original goal, to retain and preserve existing knowledge extracted by its component models from distinct streams of a text.

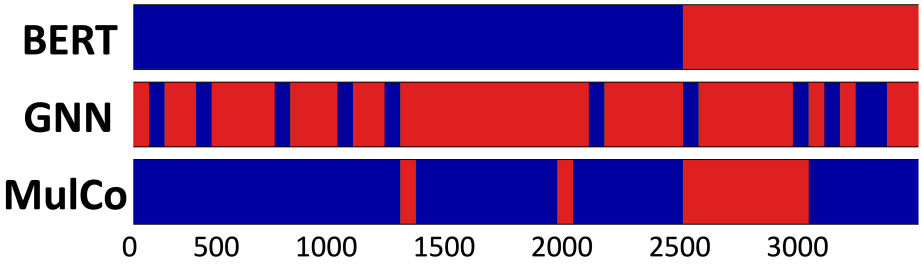

Figure 62: Comparison of distinct event-pairs captured by BERT alone, GNN alone, and the final MulCo model.

Distance	**Short**	**Long**
TDD-Auto Number of Test	877	2759
Unique Prediction		
BERT	270	795
GNN	22	83
MulCo Prediction in Unique Prediction		
BERT	234 (87%)	710 (89%)
GNN	12 (54%)	53 (64%)

Table 26: Number and percentage of MulCo predictions that carry over from BERT and GNN unique prediction.

7.3 Discussion

Traditionally, the success of TEO/ETRE tools are measured using F1 against standard benchmarks. However, this dissertation argues the standard method actually hinders the creation of models which learn meaningful features about temporal ordering. Much of positive performance by existing models can be explained by strong biases towards majority classification labels; the existence of labels like *vague* (discussed in more detail in Chapter 4) further incentivize models to avoid making predictions rather than learn the feature sets which can explain temporal ordering. The MulCo model did improve upon the SOTA of the time using basic F1 metrics, but it was qualitative analysis of which pairs the model succeeded at which informed this dissertation.

For example, one experiment had examined the confusion matrix of the final MulCo predictions. (The exact table cannot be replicated here, as the model architecture has been adjusted since that experiment was run.) In this confusion matrix, MulCo had demonstrated an over-reliance on the majority classification label of $before$–a similar problem to other models over-predicting for *vague*. This dissertation takes the stance that chronological ordering is the default method of communicating temporal information in human text, but it is essential that a model is able to recognize cases which deviate from this default. This represents an area in which future models can improve on MulCo performance, and highlights the limitations of purely quantitative analysis of model performance.

7.4 Contributions

The fundamental assertion of the dissertation is that information about time and temporal reasoning within a text is conveyed using different streams of the text itself. Much of NLP approaches 'reasoning' with the assumption that more sophisticated context-based language models can eventually mimic the types of long-distance knowledge that allow humans to reason about text. However, in TEO/ETRE, fusions of short- and long-distance approaches outperform these single-stream language models on their own. MulCo in particular is built on design principles of knowledge retention and synthesis, which makes it ideal to test the role of temporal knowledge layers on reasoning task performance.

The model had improved on SOTA for TEO/ETRE at the time of publication in three of four benchmark datasets, showcasing a particular improvement on sets featuring long-distance pairs. As these are often the most challenging pairs for neural models to predict for, the architecture is well-suited to additional improvement. Significantly, it achieves its performance by synthesizing knowledge across different layers of text. In results and analysis, the ways in which this knowledge is integrated into a model matters.

MulCo's performance informed the development of the next phase of the dissertation: the TempR-MInt model experimentation. The architecture of MulCo is complex, and information will be distilled with different weight depending on where it is introduced to that architecture. Therefore, the TempR-MInt work focused on gaps in the existing BERT/GNN fused approach, and supplied NarraType as a possible solutions for these gaps. These experiments will be described in more detail in the next chapter.

8 TempR-MInt Experimentation Pipeline

8.1 Introduction

In the **Temp**oral **R**easoning **M**ulti-Scale **Int**ention experimentation pipeline (or **TempR-MInt**) the work returns to our remaining unanswered research questions:

- **Q5. In what ways can authorial intent be incorporated explicitly into existing TEO/ETRE model pipelines?**
- **Q6. Does incorporating authorial intent into TEO/ETRE models improve performance?**

To answer these questions, the work applies a variety of quantitative and statistical techniques to the **IlNeS** corpus developed in Chapter 5 and analyze the results. Finally, we run the **MulCo** TEO/ETRE prediction model (Yao et al., 2024) with the inclusion of NarraType authorial intent information to measure performance.

8.1.1 Motivation

The critical axiom that underpins our framework for timelines is that any text's timeline consists of two sequences: the **textual timeline** (T_{txt}) and the **true chronological timeline** (T_{chr}). These two sequences (by definition) contain the same events, but possess slightly different properties. The inherent distinctions between textual and chronological timelines are as follows:

1. In a chronological timeline, multiple events may occur *simultaneously*. In text, it is impossible for two words or word phrases to occupy the same position, and therefore multiple simultaneous events must be conveyed one at a time.
2. Events in a chronological timeline may have a long duration. It is therefore possible for events to "include" one another chronologically. In text, an event mention represents the entirely of an event (specific phrases about onset or completion of an action are often classed as separate events themselves). It is therefore not possible for a textual event mention to appear "around" another event. To account for this, when comparing chronological to textual timelines we count only the starting point of a given long event.

Beyond these inherent limitations of the format, textual timelines often behave differently from the corresponding true timeline. In TEO/ETRE, the goal is for the model to correctly predict the relationship between Events A and B in a text (where A is mentioned before B in the textual timeline): for most models, the majority predicted label will be Event A is $before$ Event B, and in general trend towards the labels which represent chronological ordering (the others being *includes* and *simultaneous*). However, this reliance on a majority label can lead to the model learning less about textual cues which might generalize better to new data. As this dissertation notes, temporal deviations are *motivated* where they appear in text. The right metrics allow researchers to surface these deviations, helping to build an understand of the intentions behind the overall timeline sequence.

8.1.2 Experimental Design

An essential insight of the MulCo experimental analysis was the impact of prior class bias on TEO/ETRE predictive models. Prior benchmarks in the field have high proportions of *vague* classifications which may bias a model towards not predicting for the order of two pairs,

and after *vague* the most common label given to two text-ordered event mentions[83] is $before$. This means that, when texts do make meaningful predictions about the temporal ordering of events, their predictive output does not significantly differ from a simple majority classifier. The architectures constructed to capture structural information (see Section 3.1.3) improve on this majority classifier by capturing *exceptions* to this rule–note that this approach does match the main axiom of the dissertation (that humans default to simple chronological ordering of events when communicating time).

However, it is a limited approach to the TEO/ETRE question, and it becomes difficult in analysis to ensure that correct predictions of the $before$ label are due to temporal features instead of simple majority guesses. To reduce this risk in our analysis, the IlNeS narrative randomly scrambles the order of event-pairs so that only half reflect their in-text ordering (shown in Table 27). The 'scrambling' process and motivations behind it are discussed further in Section 8.2.5.

	IlNeS (before)	**IlNeS (after)**
Before	.75	.41
After	.08	.42
Simultaneous	.05	.05
Includes	.10	.06
Is Included	.02	.06

Table 27: Pairs by gold-standard label before and after random scrambling.

By reducing the majority impact of a single classification label, the model is forced to actively recognize chronological pairs when they appear in training data, instead of assuming their existence as default. This does align closer to distributions from TimeBank ($before$ and $after$ have similar frequencies while *simultaneous*, *includes*, and *isincluded* appear at near-identical frequencies), but there are distinct distributions of labels per dataset.

8.2 Experimental Methodology

In TempR-MInt experiments, assume this trend towards majority classification has been overcome. To test the dissertation research questions **Can integrating authorial intention information into existing TEO/ETRE pipelines improve performance?** the work modifies the initial MulCo architecture in three new ways:

1. Simple insertion of our four narrative types from IlNeS partitions.
2. Integration of LIWC feature data, which correlates with narrative types.
3. Multi-task training using narrative type labeling to accentuate TEO/ETRE prediction.

The work also explores the differences between training *across IlNeS* and training *per NarraType partition*. The latter method will always be an infeasible approach to broader TEO/ETRE modeling (as it essentially builds multiple models for general TEO/ETRE) but may reveal useful insights for the three proposed architecture variations.

[83]That is to say, within the pair (E_1, E_2), E_1 appears in the text before E_2.

8.2.1 Train by Partition

Analysis of the corpus (see Chapter 5) demonstrates that distinct NarraType partitions behave differently for TEO/ETRE. Experimentation in this section intends to reveal if this distinct behavior influences TEO/ETRE model, and therefore evaluate the utility of NarraType as a distinguishing feature for this task.

One simple test takes the MulCo baseline configuration and builds train, validation, and test sets using only texts within a single NarraType partition. This test proposes a trade-off: the training data is significantly reduced (all runs of the model have fewer than 2000 training samples), but there may be less noise by focusing on the behavior of a single NarraType. This experiment therefore expects some drop in performance compared to the full-sized model benchmark, which may or may not be overcome by focusing on a single partition.

8.2.2 Simple Type Insertion

To test the hypothesis that narrative type information can aid in TEO/ETRE prediction, experiments begin with a simple insertion of narrative type annotations into the combined BERT/GNN embeddings[84]. Given the existing MulCo architecture, there are three segments of the pipeline (shown in Figure 63) where narrative type information can be inserted.

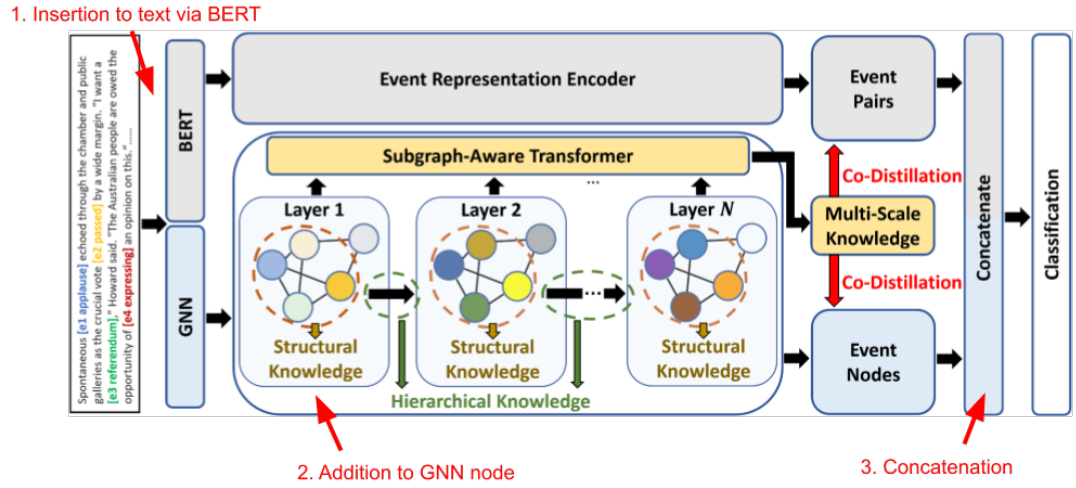

Figure 63: Insertions of narrative type information to existing architecture.

Insertion spots are described in more detail below:

1. Information can be "injected" into a BERT embedding by way of altering the original text. In this case, TempR-MInt inserted the label of the narrative type into sentences of the original document. Insertions were limited to every third sentence[85] and formatted insertions as parentheticals, to minimize the possibility that these insertions might alter the meaning of the surrounding sentence, like so:

[84]Similar to how Daumé III (2007)'s FEDA approach could expand model performance to new domains by simple augmentation of feature space.

[85]Due to the context window size selected for BERT embeddings, this ensures narrative type information is included in every potential event-pair.

"It all started by getting tired all the time, constant thirst, blurry vision and frequent infections (Factual)."

2. Though the GNNs encode structural information through node edges, the embeddings themselves contain node-specific information. TempR-MInt concatenated flags for the narrative types to these embeddings, experimenting with one-hot vectors and with integer values.

3. In the post-distillation step, representational values are concatenated before final classification. By adding a final vector for concatenation representing the simple narrative type feature, TempR-MInt can add narrative information in a format that preserves it across training.

8.2.3 LIWC Features

To produce a better understanding of the linguistic and rhetorical elements of our IlNeS corpus, TempR-MInt looks at the Linguistic Inquiry and Word Count 2022 feature set (LIWC-22) (Boyd et al., 2022) for each of our narrative type partitions. Using this data, we seek to determine if LIWC features correlate well enough with both our narrative typology and timeline ordering to serve as useful intermediate features for our model.

LIWC-22 is a software package which includes over 100 dictionaries of word types, designed with an eye towards capturing human social and psychological states known to be reflected in natural language (Boyd et al., 2021). This allows it to associate words in ways the more granular BERT and RoBERTa transformers may not, based on the surrounding context. One benefit of this approach is that LIWC features can be extracted from a text without direct knowledge of that text's narrative intention.

The LIWC-22 dictionary framework is hierarchical. Methodologically, the dissertation is particularly interested in the **Affect** and **Perception** super-dictionaries, which may reflect fundamentally distinct perspectives present in each NarraType. Additionally, the super-dictionary **Social**, **Culture**, **Lifestyle**, **Physical**, and **Motives** each had potential salience for the work. (In addition, the dictionaries **Need** and **Want** within the **States** super-dictionary could provide interesting information per NarraType.)

Multiple modes of analysis are performed using LIWC: first, features from the full LIWC-22 dictionary set for which there exist significant differences between documents of distinct NarraTypes are identified. The work limits search space to p-values in the range $p < .01$ to reduce the risk of random chance affecting results. Next, these features are used to build a discriminant analysis model to predict for NarraType (to more rigorously test their correlation). Finally, LIWC-22 features are extracted for event-pair context windows and added to MulCo architecture.

8.2.4 Multi-Task Training

The MulCo model is designed to incorporate multiple streams of information through knowledge co-distillation. This process adds to the complexity of the architecture, and thus far our effort to integrate narrative knowledge have not shown significant promise. Before we add narrative type information with co-distillation, we test its value using simpler multi-task training. The multi-task setup takes the same input as for TEO/ETRE prediction, along with labels for the narrative type of the model. The MulCo pipeline is trained for both tasks, with development holdout data and final testing only concerned with TEO/ETRE (pipeline shown in Figure 64).

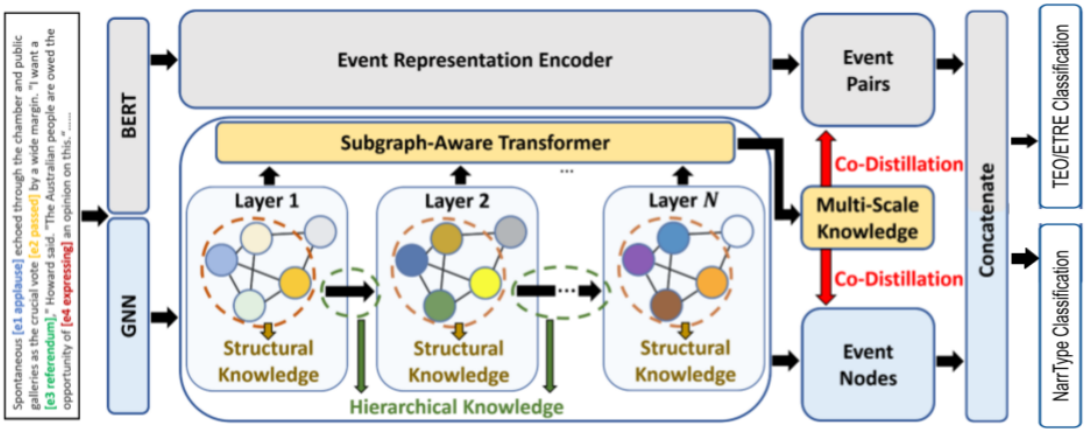

Figure 64: Addition of secondary task to MulCo model.

8.2.5 Baseline Distributions

The IlNeS corpus covers 106 participant self-reported testimonials. Of these, 94 contain annotated timelines. In total, there are 7181 event-pairs across the entire corpus. This puts it on par with datasets TimeBankDense and TDDiscourse-Manual (see Table 28), and uses 5 TEO/ETRE labels.

Dataset	Size	Labels
MATRES	276	4
TB-Dense	6088	6
TDDMan	6150	5
IlNeS	**7181**	**5**
TDDAuto	41302	5

Table 28: Comparison of corpus sizes for TEO/ETRE.

To produce a baseline for performance of MulCo on IlNeS, the TempR-MInt work takes three initial steps: 1) it overcomes a tendency in the model toward majority bias; 2) it adapts MulCo to be sensitive to the text order of event-pairs; 3) it tunes parameters for the new dataset. This section discusses these steps.

Overcoming Majority Bias:

The IlNeS dataset exhibits a strong bias towards the *before* label. However, this distribution does change with NarraType (see Figure 65).

As was indicated in Section 5.3.2, the Factual NarraType showcases the highest proportion of ordered *before* labels. Using sequence similarity, it had been noted that Persuasive and Emotional showcased the highest proportion of timeline deviations. It is interesting to note here that Persuasive NarraTypes have the smallest proportion of *before*-labeled pairs but are most distinct in the proportion of *includes* labels, which technically reflect a chronological timeline. Emotional NarraTypes have the closest proportion of *before* labels to the dataset average, but have the highest proportion of *after* and collective non-chronological labels overall.

To answer the research question: **The TEO/ETRE labels captured per NarraType have**

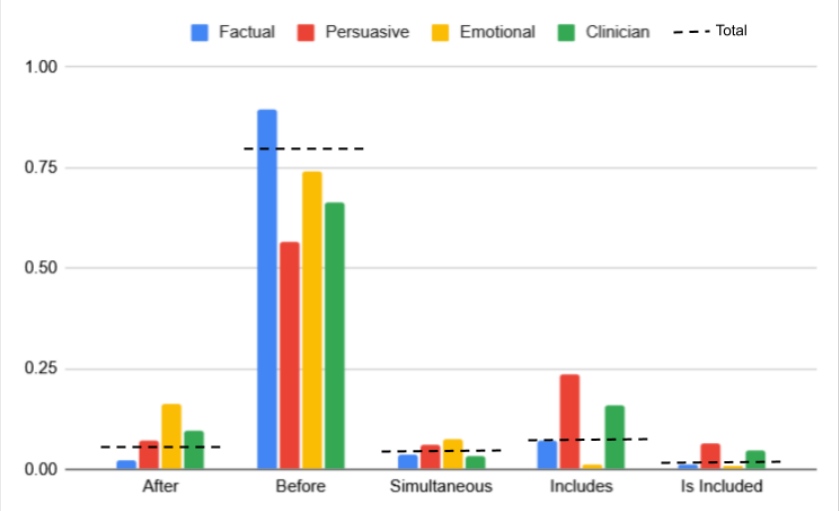

Figure 65: The proportions of TEO/ETRE labels per NarraType, including across the whole corpus (marked by black dotted line).

significant range, and the ideal bias towards the majority differs from the bias that would be learned by undifferentiated training on all IlNeS pairs. Further, simple experimentation reveals that the default behavior of the MulCo model when trained on this dataset is indistinguishable from simple majority classification predicting *before*.

Simple majority classification will produce modest F1 performance for the IlNeS test data. This value, and the values per NarraType, are displayed in Table 29.

NarraType	F1
Factual	80.2
Persuasive	40.8
Emotional	63.2
Clinician	53.1
TOTAL	64.3

Table 29: Majority classifier performance per NarraType.

These scores correlate with the overall proportion of *before* labels in the dataset, with Factual seeing highest benefit from the majority classifier. However, majority classification is not useful for true TEO/ETRE prediction, where we expect to see minority labels correctly predicted. A challenge in adapting MulCo for IlNeS was that the initial settings which were successful for TimeBank-based corpora did not differ from this simple majority class prediction when applied to IlNeS. To leverage the advanced architecture of the MulCo model for effective TEO/ETRE, TempR-MInt experiments had to demonstrate that the model was learning to perform real temporal reasoning, rather than simply defaulting to a majority label.

There were two methods used by the 'baseline' TempR-MInt implementation to overcome the tendency towards majority classification in the model: 1) dataset randomization and 2) label smoothing.

In the initial input data of IlNeS, $Rel(E_i, E_j) = label$ for all E_i, E_j in each corpus document such that the following criteria are met: a) E_i and E_j are distinct events, b) E_i appears before E_j in the original document text. Data randomization selects half of these pairs at random and instead provides the dataset with $Rel(E_j, E_i) = Reverse(label)$. This ensures that the overwhelming bias towards the $before$ label is split between $before$ and $after$. This randomized dataset is what will be referred to as **IlNeS** through the remainder of the dissertation and the raw pairs as **IlNeS***. Figure 66 shows the new distribution of labels after randomization; see Appendix 10.4.1 for more details on this final version of IlNeS used in the study.

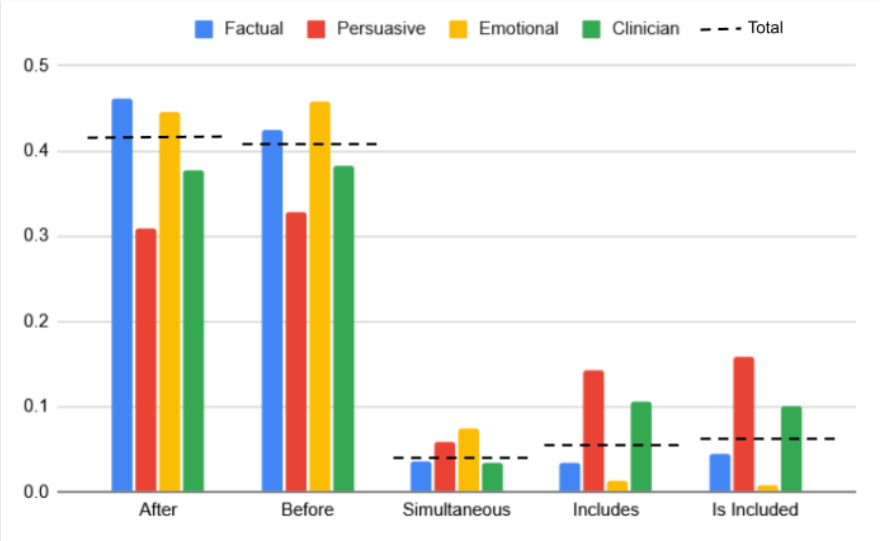

Figure 66: The proportions of TEO/ETRE labels per NarraType after pair-order randomization, including across the whole corpus (marked by black dotted line).

The new distribution of the dataset is more even, and the labels $before$ and $after$ appear with near-identical frequency per NarraType. However, these two labels dominate the dataset; initial TempR-MInt experiments on this version of IlNeS found the model predicted labels of $before$ and $after$ for all pairs in the dataset. It is important for TEO/ETRE that models be able to predict for less common label types (especially for NarraTypes like Persuasive, where these form a significant proportion of pairs). Therefore, TempR-MInt next introduced **label smoothing** to ensure the model understood cases of temporal reasoning which corresponded to these less-common labels.

Label smoothing is a feature of PyTorch (Paszke et al., 2019) which allows cross entropy loss to mix its targets between the ground truth and a uniform distribution using weights between 0.0 (full ground truth) and 1.0 (uniform). It functions as a parameter passed to the loss function. Early experimentation found that for this configuration of IlNeS, significant label smoothing weights were needed to see any of the three less-common labels in prediction; the optimal value for this dataset was determined to be 0.8–the model has access to the prior class biases of the ground-truth data, but prioritized learning information about the labels *simultaneous*, *includes*, and *isincluded*.

Text-Order Sensitivity:

Before the MulCo model can be used as a baseline to test the utility of the NarraType variable, one necessary adjustment needs to be made to how the model deals with the *context window* fed to BERT encoders to extract bi-directional context data.

In MulCo, the context window is defined as the area of document text between each event mention in the pair (inclusive of the mentions). Originally, MulCo was given data only from datasets ordered using the same criteria as IlNeS*–for each (E_i, E_j), E_i appeared before E_j in the original document text. In all cases where MulCo might be given some pair (E_j, E_i), it extracted the equivalent context window of (E_i, E_j) with no features to differentiate the two versions of the same context window. This version of MulCo achieved similar performance on IlNeS to a model which behaved like a 2-label majority classifier, making random guesses between the labels $before$ and $after$[86].

To fix this, the TempR-MInt work introduced a set of features to the final concatenation step, which encoded the Boolean value $TextOrder(E_x, E_y) = true$ iff E_x appears before E_y in original document text. There is significant improvement in MulCo performance with this simple addition, which we call MulCo-Sensitive (as it is sensitive to the textual order of events) or **MulCoS**. This model forms this dissertation's comparative benchmark for testing the NarraType variable.

Model + Dataset	**F1**
Majority classifier + IlNeS	24.7
2-label random prediction + IlNeS	41.4
Majority classifier + IlNeS*	64.3
MulCo + IlNeS	45.2
MulCoS + IlNeS	**66.1**

Table 30: IlNeS benchmarks without NarraType features.

The randomized IlNeS dataset presents a higher level of challenge than the raw IlNeS* and relies significantly less on majority bias towards a single classifier (simple majority achieves approximately 39.6 F1 points higher for IlNeS* than IlNeS). MulCo does improve upon a basic heuristic approach for IlNeS by 3.8 F1 points, and adding sensitivity to textual order improves F1 by 19.7 points. It is noteworthy that this benchmark, despite using the more challenging IlNeS dataset, now achieves improved performance compared to the majority classifier on simpler IlNeS* and correctly identifies more types of labels (see Appendix 10.4.1 for the confusion matrix Table 59).

Parameter Tuning:

The next step outlines the ideal parameters for MulCo for the IlNeS dataset. MulCo's baseline implementation identifies distinct hyperparameters per dataset. The IlNeS benchmark uses these hyperparameter configurations as a guide, finding that IlNeS behaves most like TDDiscourse-Manual. Initial ablative studies found the parameters in Table 31 function best to build the baseline. Though MulCo performs best with TDDiscourse-Manual after only 5 epochs (with further epochs leading to over-training), experimentation found improvement for IlNes with number of epochs increased to 8. The similarity between performance for IlNeS and TDD-Manual makes intuitive sense; the datasets are near the same size and both contain a mixture of short- and long-distance event-pairs.

[86] See Appendix 10.4.1 for full details.

Parameter	Setting
BERT dimension	768
GNN dimension	256
RGAT attention dimension	256
SAT dimension	768
Embedding space for cl	2048
K-hops	1
Number of GNN layers	2
Dropout	0.1
Learning Rate	1e-5
Label smoothing	0.8
Batch size	16
Temperature in CL	0.1
Temperature in KD	0.1
Random Seed	2513
Epochs	8

Table 31: Optimal hyperparameters for MulCo IlNES baseline.

F1-scores for the new MulCo implementations are achieved using 5-fold cross-validation. Documents per fold are selected at random with the exception that each fold contains an even distribution of Factual, Persuasive, Emotional, and Clinician narratives. This ensures scores are robust and results avoid confounding factors like the size or narrative type of documents assigned to training.

Dataset	F1
MATRES	84.4
TDD-Auto	77.1
TimeBankDense	67.5
IlNeS	**66.1**
TDD-Man	55.1

Table 32: MulCo performance on existing datasets.

The starting baseline for IlNeS (using MulCoS) is compared against prior performance on TimeBank datasets in Table 32. MulCoS-IlNeS nears performance on the TimeBankDense dataset despite the switch between domains, suggesting the dataset presents some challenge to TEO/ETRE but is not as difficult as long-distance TDDiscourse-Manual.

8.3 Experimental Results

Implementation of all experiments (unless otherwise stated) uses the MulCoS order-sensitive model, with label smoothing parameters, and the IlNeS randomized dataset. 5-fold cross-validation is used to ensure robust measurement.

8.3.1 Training by Partition

Table 33 shows results for models trained, developed, and tested using only documents from a distinct NarraType, to test the theory that the model would naturally adjust to the correct distributions of labels per NarraType partition.

NarraType	Majority	Partition
Factual	80.2	70.0
Persuasive	40.8	51.6
Emotional	63.2	66.0
Clinician	53.1	48.9
Total	64.3	64.0

Table 33: F1 scores of partition-only training by NarraType.

Prediction on Factual NarraType documents measurably benefits from isolated partition training[87], outperforming the MulCoS IlNeS benchmark (which achieves an F1 of **66.1**) by 4 F1 points. However, this implementation under-performs compared to its partition level majority classification model–Factual data contains the highest proportion of pairs whose chronological order matches the textual, and the label smoothing hyperparameter that is most effective for the corpus overall likely has a negative effect on Factual partitions.

Emotional NarraType documents improve on their majority classification model and perform on par with the MulCoS baseline with partition training. This suggests only small or negligible benefits from isolated partition training for this NarraType. The Persuasive NarraType shows strong improvement on its majority classification model, though it significantly under-performs compared to the full-data benchmark. This suggests that isolating Persuasive documents from the surrounding datasets does help the model to identify useful features for this NarraType, but that this intention type is challenging to model overall.

Clinician NarraType documents under-perform compared to its own majority classification model and the MulCoS benchmark. This makes some intuitive sense; observation of Clinician documents (discussed throughout Chapter 5) showed that behavior within this partition could be described as a mixture of other NarraTypes. Therefore, isolating this dataset only reduces the model's available training data while providing no benefit from a tightened scope.

> *(Quick experiments show that, despite some correlation in behavior between Clinician and Emotional NarraTypes, training the model on both together does* ***not*** *improve performance beyond what is expected by averaging their prediction results.)*

These results suggest that, overall, there is some situational benefit to separating documents by NarraType for TEO/ETRE modeling. However, certain partition types are better predicted by a model which has access to data on other NarraType texts. Further, there is an unknown drop in performance from reduced training set size associated with this experimental method.

8.3.2 Simple Type Insertion

Simple insertion of NarraType information (shown in Table 34) produces comparable performance and minor improvement on the baseline. It is not clear that these improvements (.6 F1 points in the largest case) represent a significant improvement rather than simple random variation. Although 5-fold cross-validation makes results more robust, it cannot fully eliminate

[87]Note that the Factual NarraType includes a single document containing 2000 pairs on its own. The basic architecture of MulCo splits data into train/val/test sets on the document level. The test of the Factual partition therefore saw a significant drop for the case when this document was sorted into testing, which left the training data set significantly smaller than test data. To ensure a similar train-to-test ratio as for prior benchmarks, the Factual partition work split into additional validation folds to ensure half of this large document remained in training at all times.

Model	F1 (short pairs)	F1 (long pairs)	F1 (all pairs)
Non-NarraType Benchmark	51.9	**69.8**	66.1
BERT insertion	50.2	68.9	64.9
GNN insertion	53.9	69.3	66.2
Concat insertion	53.6	68.9	65.9
BERT + GNN insertion	**54.9**	**69.8**	**66.7**
BERT + Concat insertion	54.0	69.2	66.0
GNN + Concat insertion	53.1	69.4	66.1

Table 34: Impact of type insertions on MulCo framework.

the potential for random chance to create perturbations in model output. The most successful implementation of the model inserts NarraType information at both the BERT and GNN steps.

Output for MulCoS divides F1 between short-distance and long-distance event-pairs; though insertion of NarraType rarely produced significant improvements to overall model performance, it does consistently improve short-distance pair output (3 F1 points improvement in the best case). This is likely to reflect a true difference caused by the model, and suggests that direct NarraType incorporation has more impact on nearby event mentions within a text.

8.3.3 LIWC Features

To examine the differences in LIWC features across NarraType, LIWC features are collected per document and sorted by NarraType. The work uses two-sample Kolmogorov–Smirnov testing to compare the distributions for each NarraType pair, and highlights only the cases with $p < .01$. Even with this cutoff, it remains possible that the variation between certain features across narrative types is due to random chance. Specific examples, including their p-values, are discussed in more detail within this section.

Topical Differences:

TempR-MInt expects some natural variations in topic as a natural consequence of illness narrative sub-types that are not otherwise useful to our work. For example, the Persuasive subtype shows the highest proportion of Money words (from super-dictionary **Lifestyle**) compared to the other categories. This makes sense (as the Persuasive prompt asks participants to imagine their audience speaking to someone who may allocate funds to research) but does not tell us much about the rhetorical distinctions between subtypes.

In Table 35, we show which NarraType partitions demonstrate high- and low-value distributions for topic-related LIWC categories. Only categories for which a p-value $< .01$ are shown here; full means and p-values may be found in Appendix 10.4 under Table 60.

Despite the purpose of IlNeS corpus being to collect testimonials about illness, there are statistically-significant variations in the $Health/Illness$ and similar categories. In particular:

- Factual narratives were *never* characterized by a high proportion of topical language compared to other NarraTypes.
- Persuasive narratives had the highest number of frequent topics, including distinct topics like $Money$ and $Work$.
- Emotional narratives have a low frequency of language relating to most topics, *except* for social topics ($SocBehav/SocRefs$) and the $Want$ category.

Partition	High-Value Categories	Low-Value Categories
Factual		Lifestyle, Money, SocBehav, Want, Work
Persuasive	Health, Illness, Lifestyle, Money, Physical, Work	SocRefs
Emotional	SocBehav, SocRefs, Want	Health, Illness, Lifestyle, Money, Physical, Work
Clinician	Health, Illness, Physical	Lifestyle, Money, SocBehav, SocRefs

Table 35: Topic-related LIWC categories which appear in different NarraTypes at high and low frequencies.

- Clinical narratives have a high frequency of topical language most directly related to health and physical function, and low frequencies of many other topics.

Rhetorical Differences:

There are clear and significant differences in style per partition. This information has obvious utility for modeling and analysis and is important to identify here. A summarization of feature frequencies can be found in Table 36 and the full values in the appendix under Table 61.

Partition	High-Value Categories	Low-Value Categories
Factual	Negate	Analytic, BigWords, Clout, Prosocial, Tone, TonePos
Persuasive	Analytic, BigWords, Clout, Prosocial	Allure, Assent, Authentic, Conversation, Linguistic
Emotional	Allure, Assent, Conversation, Clout, Insight, Linguistic, Prosocial, Social, Tone, TonePos	Analytic, BigWords
Clinician	Authentic	Assent, BigWords, Clout, Conversation, Insight, Negate, Social, TonePos

Table 36: Rhetorical LIWC categories which appear in different NarraTypes at high and low frequencies.

Note the following differences in rhetoric:

- Factual narratives are given few rhetorical flairs, with the exception of a high frequency of *negating* language. This can be explained by participants using Factual NarraTypes to correct misinformation, which requires negating the common belief.
- Despite the purpose of a factual narrative being to educate an acquaintance, Factual texts were less likely to use language which positioned the speaker as particularly knowledgeable (*Analytic*, *BigWords*) or authoritative (*Clout*). Various social language was also less likely to appear here.
- Persuasive narratives have a high frequency of rhetorical types which correspond with appearing knowledgeable (*Analytic*, *BigWords*) and authoritative (*Clout*). There ap-

pear to be overtures to a shared common good (*Prosocial*), which may be seen as good persuasive strategy when discussing illness research in specific.

- Persuasive narratives were least likely to contain *Authentic* language.
- Emotional narratives have the largest distribution of rhetorical language types, with positive tones (*Assent* and *Tone/TonePos*), social rhetorical skills (*Allure*, *Conversation*, *Prosocial*, *Social*), as well as emotional categories (*Insight*).
- Emotional narrative were less likely to use scientific language (*Analytic*, *BigWords*).
- Clinical narratives only contain high proportions of *Authentic* rhetorical language. The clinical narratives featured more language meant to express authenticity compared to every other NarraType.
- Clinical narratives have low proportions of some types of scientific language (*BigWords*) and many social rhetorical types (*Conversation*, *Insight*, *Social*, *TonePos*).

This dissertation highlights the behavior of the *Authentic* category in Clinician NarraType, which has interesting implications for patient-clinician communication. It suggests that patients do on average seek honest communication with new providers[88].

Time:

Due to the emphasis in this dissertation on elements of time, it is worthwhile to note cases where temporal language differs across narrative type. A summarization of feature frequencies can be found in Table 37 and the full values in the appendix under Table 62.

Partition	**High-Value Categories**	**Low-Value Categories**
Factual	FocusFuture	
Persuasive		
Emotional	FocusFuture, FocusPresent	Time
Clinician	Time	FocusFuture, FocusPresent

Table 37: Time-related LIWC categories which appear in different NarraTypes at high and low frequencies.

The results show the following:

- Emotional narratives have a greater focus on present and future events than other NarraTypes.
- Clinical narratives do not often focus on present and future events, suggesting the highest priority for clinical narratives (which simulate initial discussions with a general practitioner) is past medical history.
- Non-specific temporal language is used most in clinical narratives and least in emotional narratives.

Other:

The work in illness narrative by Bury (2001) suggests a trend among authors of such narratives to utilize moral arguments in the retelling of their story. This informed construction of the Persuasive narrative sub-type, and this work hypothesized that these moral arguments would

[88]The Clinician prompt specifies a first meeting with a new GP.

be more prevalent among Persuasive narratives than others. There was evidence of Bury's moral narrative across all illness narratives in Chapter 5, but the work did not observe notable correlations with any narrative type. The LIWC category "Moral" gave TempR-MInt a chance to examine that hypothesis with statistical data, and the work found the following: **there is no statistically-significant difference in moral language across any of the narrative survey subtypes**[89]**.** Participants will make moral arguments in their illness narrative at similar rates regardless of the imagined audience.

Linear Discrimination:

The work demonstrates the correlation of LIWC features with narrative type by producing a simple linear discrimination model. This model predicts for document narrative type with 81% overall accuracy. The confusion matrix for prediction is shown in Table 38:

	True Values			
	Factual	Persuasive	Emotional	Clinician
Fact. (Predicted)	22	4	2	3
Pers. (Predicted)	1	20	0	1
Emot. (Predicted)	0	1	21	0
Clin. (Predicted)	4	2	2	22
Accuracy (%)	81.5	74.1	84.0	84.6

Table 38: Confusion matrix of discriminant predictive modeling using LIWC.

This project explores the following options to incorporate these features into MulCo, similar to simple type insertion:

1. Insertion at the GNN node. By appending the LIWC features as floats within the GNN node structure, LIWC features can be integrated to the structural half of MulCo.
2. Insertion at the concatenation step.

LIWC features can be extracted on the document level or on the sentence-level, though the latter takes more time to extract and test.

Model	F1 (short pairs)	F1 (long pairs)	F1 (all pairs)
Non-NarraType Benchmark	51.9	69.8	66.1
Sentence-level GNN insert	51.0	68.2	64.8
Doc-level GNN insert	**53.5**	**70.1**	**66.8**
Doc-level concat insert (features)	53.3	69.7	66.4

Table 39: Impact of LIWC insertions on MulCo framework.

Performance using these LIWC insertions is shown in Table 39. Like with simple type insertion, the small improvements are not outside the bounds of ordinary variation from the baseline. In particular, results for GNN insertion became less accurate when LIWC features were distinct for nodes across a single document. As GNNs better learn structural information when attributes differ across edges, this suggests that more meaningful LIWC features confound a GNN, while the model can ignore document-level attributes. Therefore, 'best' performance for this insertion type may reflect a chance variation from baseline performance.

[89]The lowest p-value differentiation between two NarraTypes was .189 for this type of language.

Multi-Task Training:

Early experiments with multi-task training (shown in Table 40) suggested the approach had little potential value. Numbers are recorded here for completion's sake. Note that, due to the computational cost of multi-task training, initial tests had been performed on a smaller subset of IlNeS. The scores for this comparative baseline are listed in the table, and do not match the scores for the full IlNeS benchmark.

Model	F1 (short pairs)	F1 (long pairs)	F1 (all pairs)
Non-NarraType (small)	**34.5**	**51.3**	**44.4**
Multi-task	25.5	36.6	34.5

Table 40: Impact of multi-task training on MulCo framework, compared against a small IlNeS baseline.

This early experiment of multi-task caused performance to drop by at least 10 F1 points across all categories. This suggested that multi-task was not a promising experiment to attempt on the larger IlNeS corpus. This makes some intuitive sense, as multi-task reflects a simple method of aggregating multiple streams of knowledge in a single model and is often less efficient than other methods of sharing knowledge. MulCo in its initial formulation uses a more sophisticated method of knowledge co-distillation. With multi-task experimentation, our goal was to demonstrate whether NarraType information improved the model in any respect, before moving on to the more complex knowledge transfer.

8.4 Discussion

The TempR-MInt project responds to the research hypothesis **The intention with which an author chooses to write a text will change the order that events are presented within the text**. It answers the following research question: **Can integrating authorial intention information into existing TEO/ETRE pipelines improve performance?**

Though there are some correlations between NarraType and chronology, it is difficult to incorporate NarraType data successfully into the MulCo pipeline. The dissertation sees the most promise for **short-distance** pairs, achieving up to 3 F1 points of improvement. This which may suggest that NarraType is most effective on nearer event mentions, contrary to expectations that NarraType is a function of the *structure* of the text. A strong improvement was made to MulCo in incorporating *text order sensitivity*, but this is separate to the fundamental hypothesis that narrative intention will contribute to improved TEO/ETRE.

The advancements achieved by the TempR-MInt experimental project may show promise in the future. However, on the full, comprehensive IlNeS dataset, changes in behavior due to intention information have not shown meaningful differences thus far from random chance.

8.5 Contributions

TempR-MInt work builds on MulCo by presenting it with a new stream of knowledge: narrative intention. As MulCo functions by effectively synthesizing knowledge across layers of text, it was ideal to test the role of temporal knowledge layers on reasoning task performance. TempR-MInt evaluated the impact of simple NarraType data insertion, LIWC features as narrative-rhetorical proxies, and multi-task training approaches.

In these results and analysis, it is clear the ways in which new streams of information are integrated into a model matter. For many cases, the introduction of NarraType information could distract the model from its primary goal of TEO/ETRE prediction. This is especially true when attempting to align embeddings for both TEO/ETRE and NarraType, where simple multi-task training causes an immediate and dramatic drop. Other studies using MulCo for model alignment note its knowledge distillation may not see strong results in cases where knowledge layers do not capture complementary signals (Wu et al., 2025). Chapter 5 shows correlation between NarraType and trends in TEO/ETRE per document, which may suggest that there are other ways to present this knowledge layer to MulCo to leverage that correlation. The unique knowledge co-distillation method produced by the original MulCo work requires precise integration of new layers of data, but generally provides performance boosts for layered input tasks. Though this initial work did not show strong results for the task, future work aims to experiment further.

These experiments, and others performed on the MulCo model, do show the utility of a *qualitative*, *focused* approach on the TEO/ETRE task. By highlighting gaps in the existing BERT/GNN fused approach, and extracting streams of temporal reasoning needed to fill these gaps, ongoing work can approach the task more effectively. MulCo's initial implementation found that accounting for both short-distance context and structural, hierarchical relationships provided stronger results than a purely linear modeling approach; in the TempR-MInt experiments, the dissertation explored ways this logic could apply to other forms of temporal knowledge. Modest success informs where the model can improve in the future.

9 Conclusions

This dissertation work sought to answer if there existed observable correlations between the intended impact of a text on an audience (simplified to "authorial intent") and the temporal order of events within the text. Broadly, analysis of the novel IlNeS corpus does support this hypothesis: narratives intended to be factual are most likely to conform to a simple chronological ordering; narratives intended to persuade tend to feature temporal deviations and inverted timeline segments; and finally, narratives intended to invoke emotional connection feature temporal deviations and interleave elements from distinct segments of the timeline.

The work began from a simple intuition: if the temporal elements of text are most clearly understood by using chronological order, then there must be motivating factors which lead human authors to introduce temporal deviations to text. There must be some unconscious trade-off where the benefits to the text outweigh the associated reduction in temporal clarity. The findings from IlNeS support and expand on this general intuition. Chronological ordering *is* associated with clear communication of time, and is most common when a text's author primarily intends to use a text to educate. The other intention categories, it appears, require more complex in-text timelines. Deliberate introduction of select temporal deviations seem to improve a text's ability[90] to use a narrative to persuade a reader to a certain line of thinking, or forge emotional connection between author and reader. It should be noted that these deviations are not likely the underlying cause of a text's persuasive or emotional potential, but rather are themselves signals of a text-wide holistic process where a narrative is shaped by the author to suit a certain end.

This insight has strong potential to inform future directions in narrative, linguistic, and computational study. Temporal reasoning is a notable task within NLP, and the connection to narrative is one not currently represented by SOTA. This dissertation does note that, while useful trends within the data were identified, it remains difficult to directly integrate information about a text's narrative into existing machine learning models. To discuss the specific benefits this dissertation can provide its audience, this chapter will break the research hypothesis down into the six research questions first formulated in Chapter 1 and discuss. Then it will sum up formal contributions and future work.

9.1 Research Question Discussion

This dissertation asked six research questions, answered here with insights pulled from across the research work.

Q1. How can we effectively annotate new corpora for temporal elements, given the inherent challenges of this domain? This question is essential for NLP researchers looking to expand temporal reasoning datasets to new domains and corpora. The task presents with a high level of challenge despite its significance.

Annotators for TEO/ETRE can be grouped into three types: trained annotators specializing in time; trained annotators without temporal specialization; untrained annotators relying on human intuition. Methods for eliciting new annotated corpora depend on the annotator type; however, in general this work finds that leveraging those human intuitions of time streamlines the annotation process when annotators belong to the latter two groups.

Timelines of events are most intuitively conceptualized as holistic objects, rather than as sum totals of annotated event-pairs. Annotators in the first group (trained for temporal annotation)

[90] Or, at least, a human author's perception of the text's ability.

can understand event-pair annotation with some degree of reliability. However, even this group fails when assigned challenging pair types (like long-distance or semantically unrelated events). One method of countering this difficulty is to guide annotation with heuristics like date-time comparisons, causal semantic links, and social scripts to produce a subset of existing events. These human-annotated could be supplemented by leveraging the interdependent nature of event-pairs produce further labeling (a process which can be automated, saving time). This methodology shows reasonable success with trained annotators specializing in temporal labels.

When working with non-specialized annotators (individuals who have been trained on annotation schema, but not schema directly instructing on time), the level of challenge increases. Best results were found by framing pair annotations as elements which contributed to a single holistic timeline. This greatly improved annotation efficiency and decreased cognitive load on annotators, who found this approach significantly more intuitive.

These insights contributed to the design eventually used for IlNeS, where untrained annotators generated a chronological timeline corresponding to their own illness narrative testimonials. Given the extreme constraints of the survey format, the annotation schema had to be clear and concise, while still capturing the necessary complexity of the task. Rather than direct annotators to label individual event-pairs, the survey collected temporal information by asking for the creation of a single timeline containing all events. This annotation method is more comprehensive than competing annotation methods (which themselves require annotator trained specifically for time). Though the survey found many cases where respondents made formatting errors in their annotation, they showed an overall understanding of the conceptual basis of temporal ordering necessary for TEO/ETRE.

Q2. What can we observe about the link between temporal positioning in text and intended effect from short-form standardized narrative text? Demonstrating a link between the temporal arrangement of events in text and the text's intended effect has potential to introduce new forms of knowledge to temporal machine-learning and NLP. It validates work within literary narrative theory and provides quantitative evidence of existing hypotheses.

This dissertation does demonstrate a link between the temporal positioning of events in text and the text's narrative intent. The work divides intention into three categories: the intent to provide factual information; the intent to persuade an audience; and the intent to form an emotional connection. The dissertation also examines texts which mimic patient-clinician communication and asks if they can be linked to one of the three primary intention types.

Overall, the Factual narratives showed the least amount of temporal deviation, which fits a hypothesis that chronological ordering presents events in the clearest way to a human audience. Because the intention behind a Factual narrative is to be informative, this type of temporal positioning may be unconsciously understood by authors as the most effective. Deviations were more common in the Persuasive, Emotional, and Clinician narrative types, which have different underlying intentions. Persuasive texts were somewhat more likely to have full inversions of timelines or timeline segments, while Emotional texts showed cases of distinct topical timeline segments interleaving. At many points, Clinician texts appeared to behave similarly to Emotional texts, which may support Charon's framework of illness narrative as an effort to bridge gaps between clinician and provider. However, there remained variance within the subtype, which suggests a degree of plasticity in clinical communications reflecting individual patient priorities.

Q3. Are there certain types of narrative which exhibit temporal deviations disproportionate to prior class biases of ordering models? Biases towards a specific classifica-

tion label limit the utility of machine-learning models. This is a specific challenge within ML and this dissertation explores solutions for the task of TEO/ETRE.

Current ordering models have strong biases towards *vague* (a label without meaning along the temporal axis, but which marks certain event-pairs as difficult) and $before$ labels, which demonstrate a low level of temporal deviation within existing TEO/ETRE corpora. Contrary to initial hypothesis, the IlNeS* dataset[91] showed a higher frequency of the $before$ label compared to other corpora across all narrative intention types. This may be due to the comprehensive nature of IlNeS* compared to competing TEO/ETRE corpora; the types of pairs which are manually annotated in these frameworks may be more likely to contain rarer TEO/ETRE types. After random inversion of pair orders, the final IlNeS dataset does not have the same bias towards $before$ labels but behaves similarly for the *simultaneous*, *includes*, and *isincluded* labels.

The dissertation does find that distinct narrative types have different label distributions, and therefore models which are trained on the full IlNeS corpus may struggle to capture the distinct behavior of individual narrative types.

Q4. How is the behavior of short-form illness narratives driven by the factor of authorial intent? Illness narrative research can benefit from quantitative backing of previously-observed trends; the IlNeS corpus supports existing hypotheses in the field and enables future work to perform more rigorous analysis within the space.

Per the dissertation's findings, short-form illness narratives do demonstrate distinct behaviors based on the intentions of the author. Q2 discussed observed differences in temporal positioning in text, but there were also distinctions in rhetorical elements like tone, register, and topical focus. Prior typologies of narrative illness often focused on internal motivators of narrative (i.e. what the author of the text most needed from the chance to share their story), but providing IlNeS participants with randomly-assigned prompts appeared to change how participants approached their narrative. For example, there was some observed tendency where entering a text with factual intentions could improve a narrative's ability to "make sense" of the experience of illness. Asking participants to write as if speaking to another individual diagnosed with one's same condition evoked significant emotional outpourings, as well as advice and offers of comfort. Presenting participants with the scenario of working to persuade someone about their condition caused behavior most different from existing illness narrative taxonomies; this illuminates a gap in illness narrative theory with potential for future work.

Q5. In what ways can authorial intent be incorporated explicitly into existing TEO/ETRE model pipelines? The dissertation proposes that its methodology can be useful to ML engineers working within temporal reasoning. The TempR-MInt experiment presents possible options for future architectures building off of MulCo's knowledge distillation approach.

NarraType information from IlNeS can be inserted into a model architecture on a number of distinct layers. MulCo provides three simple insertion points: through BERT embedding, appending to GNN node features, and by direct concatenation to other base models' predictive output. NarraTypes can be encoded as a numerical vector or as a one-hot encoding, with some greater success observed with one-hot vectors. However, it is worth acknowledging the granularity of the NarraType feature. NarraType (a record of the simulated audience prompt) was collected on the document level, but it is not necessarily true that all sentences within a document fit the general pattern of this NarraType. Like in discourse theory, which splits text

[91] The corpus of raw annotations before random scrambling of pair order.

into 'discourse steps' with distinct functions that contribute to a unified end goal, individual sentences of a given NarraType text may use features of other intention types.

Therefore, NarraType information could also be encoded using LIWC feature embeddings. LIWC features across a document have significant correlation with the NarraType variable, and they can be extracted on the sentence level. This adds flexibility to the way intention information can be incorporated into the MulCo model. As with the simpler NarraType encoding, this information can be inserted at multiple points across MulCo.

Finally, multi-task prediction can be used to train a single model for both TEO/ETRE and NarraType prediction, with the idea that this approach might allow a model to surface features which correlate to NarraType, and that these may be more useful than NarraType directly.

Q6. Does incorporating authorial intent into TEO/ETRE models improve performance? This information is most useful to machine learning researchers working on TEO/ETRE and temporal reasoning more generally. Though this dissertation has presented multiple options for incorporating authorial intention into these models, it is unclear how directly useful this is for ML research.

Thus far, it is difficult to say that authorial intent improves performance beyond what could be found through random chance. There are configurations which add up to 1 F1 point to baseline MulCoS (MulCo-Sensitive, which can track the textual order of events in a pair) performance. There is observed a stronger improvement (3 F1 points) for **short-distance** even-pairs, which may suggest that narrative intention has a stronger impact on nearby pairs in a text than for distant pairs (i.e. the majority of temporal deviations introduced into even dis-ordered texts are local). The TempR-MInt experiments inform future exploration into this question, but the project has not yet found a pipeline which leverages authorial intent to improve TEO/ETRE.

9.2 Contributions

This dissertation expands existing frameworks within the fields of temporal reasoning, narrative theory, and clinical text analysis. It presents a tool for the extraction of grammatically-oriented time expression (**STAGE**) and a set of methodologies which improve annotation efficiency for TEO/ETRE and related temporal reasoning fields (each of which is scoped to a different level of annotator expertise). It demonstrates the utility of this annotation framework within IlNeS data collection, to produce a set of reliable TEO/ETRE annotations from testimonials with limited time to train participants for the task. These works represent foundational technologies for building temporal reasoning corpora, and apply insights from within other approaches to time (e.g. the logical framework of time, holistic sequence view) to the NLP space.

It contributes a publicly-accessible corpus of over 100 illness narrative testimonials **IlNeS**, linked to medical event and TEO/ETRE pairwise annotations. The survey is labeled for NarraType, which represents the simulated audience that participants were prompted to address with their testimonials. Finally, rich freeform feedback data provides direct evidence for the motivations behind various features of the illness narrative testimonial. This corpus has significant potential in a wide range of fields. It expands the NLP task of TEO/ETRE beyond the traditional domain of journalism, it provides new data within the illness narrative genre, and it features key insights about the individual patient experience with illness. The freeform feedback response provides motivation for many discourse actions and narrative techniques used throughout the testimonial data, and could therefore be of use in studies of narrative construction. Though the project was scoped specifically to extract temporal attributes from the text, post-processing could allow the raw data to be useful for many other possible studies.

Within this dissertation, specific evaluations were performed on IlNeS to extract trends in temporal behavior across NarraType. These insights suggest a correlation between the intention of a text and its temporal properties, which may be useful for narrative study and the improvement of temporal reasoning models. The methods of evaluating timelines (e.g. the use of rank metrics for quantitative measurements, grouping timelines into segments to better surface specific temporal ordering behavior) can be applied beyond the IlNeS corpus to other datasets.

Finally, the dissertation contributes an expansion of the MulCo TEO/ETRE pipeline called **TempR-MInt**, which introduces intention information at different points of the pipeline. TempR-MInt specifically overcomes a challenge common to TEO/ETRE, which is the prior case bias of models in datasets which disproportionately feature chronological event-pairs. It also presents distinct ways that new forms of knowledge (here, NarraType intention data) can be added to MulCo's multi-scale knowledge distillation framework. Though these results do not significantly improve on existing TEO/ETRE, they present many possible avenues for future work, and the methodology could be replicated for other types of textual knowledge that NLP research believes could improve temporal reasoning. The experimentation pipeline expands on this prior work, and its utility is aimed towards enabling more focused TEO/ETRE on challenging benchmarks in the future.

9.3 Future Work

In this section, the work discusses potential methods that others might use to build on this dissertation, and specific experiments to be followed up on after the dissertation's completion.

Broadly, this work lays a foundation for human-centered examination of text within both temporal NLP tasks and NLP as a whole. Though the exact mechanisms by which authorial intention may shape temporal ordering within text are diffuse, there is some correlation which can be observed between the two. Lay writers can be induced to pursue a certain style of writing when presented with a specific simulated audience, which suggests that naturally-generated writing for similar audiences also displays characteristic style. Temporal ordering, as discussed in Chapter 2, is foundational to temporal reasoning as a broader suite of complex tasks. Current ML models, which are capable of significant acts of interpretation when given a text in isolation, still struggle with these types of reasoning.

The theoretical framework central to this dissertation presents text not as an isolated object, but as a projection of reality. Like a shadow cast on a cave wall, it may require additional information for a predictive model to fully understand the ground-truth it is intended to communicate. In many cases of NLP, that information is not directly observable (consider the case of news articles in TimeBank which could have persuasive intent but which cannot be definitively marked as such), but a better understanding of human trends in writing as a whole can help models to make inferences about these types of external information.

Specific intended projects are detailed below.

Release of IlNeS data:

The IlNeS data requires minor post-processing before release to the public on GitHub. Data will be provided to future researchers in the following formats: 1) Raw texts per participant. These texts will include the survey as submitted by the participant, and therefore may include missing answers or serious formatting errors. The raw texts will be anonymized, with potentially identifiable data such as names, locations, or dates redacted. External researchers will not have access to data or metadata including the un-redacted values; 2) A CSV file with pro-

cessed data per participant. This includes the NarraType, testimonial text, ordered timeline, and freeform response; 3) The dataset as formatted for direct input to MulCo and derivative TEO/ETRE models. All participants will be identified using only anonymous, randomized ID values throughout the released corpus.

Expansion of IlNeS annotation:

Though the work in this dissertation identifies methods of surfacing notable examples of temporal deviation within a text, and provides a framework with which to analyze them from the perspective of narrative intent, its corpus analysis is limited to the specific types of data which could be collected in a study of its type. The work is limited demographically (recruiting only from adult English-speakers within United States who had some experience with illness or disability) and the choice to focus on the domain of narrative illness limits insights to that domain. To expand this type of work, data collection could be repeated using distinct domains, languages, and cultures–but to do requires collaboration with researchers who can bring expertise regarding these new areas of interest.

Further TempR-MInt Experiments:

There were multiple experiments within TempR-MInt which could not be completed within the span of the dissertation, listed here:

1. **Direct BERT Modification**: MulCo as it currently exists does not allow significant modifications to be made in the BERT embedding layer. However, this may be a more effective way to integrate NarraType information compared to text injection. These experiments would modify the MulCo pipeline to allow for this type of BERT modification.
2. **STAGE Integration**: STAGE timexes are not currently compatible with MulCo (which requires timexes to be in the specific format expected from TimeBank). To observe the impact of the STAGE timex framework on MulCo, the systems must be modified to produce and read timexes in the same format.
3. **Rhetorical Graphs**: The property of narrative intention could be conceptualized as an element of discourse. In discourse theory, texts are often broken into 'acts' or 'steps' which function together to contribute to an end goal. Similarly, it may be possible to use rhetorical structural theory (RST) graphs (an element of TIMERS that was omitted from MulCo) to better surface areas in a text which are influenced by NarraType. This could allow NarraType information to be applied selectively through a text rather than on the document level, which may better match the temporal behavior observed in this dissertation. This experiment requires implementing new RST graphs[92] for the MulCo pipeline and experimenting with NarraType.
4. **Other Multi-Task**: Experimentation surfaced correlations between NarraType and textual elements like LIWC. It is possible that performing multi-task prediction with a different secondary task will be more effective than multi-task prediction of NarraType.
5. **True Co-Distillation**: MulCo leverages distinct knowledge streams more effectively than predecessors due to its co-distillation approach. In TempR-MInt, the effects of NarraType information were measured using only simple insertion (with the idea that this was a cheaper way to identify the best approach for more complex co-distillation). Co-distillation may surface benefits from NarraType intention knowledge that are not

[92]The version used in TIMERS had not been available for use in MulCo's original implementation.

observed with more basic types of insertion; these experiments requires the construction and implementation of a more involved, three-prong model.

9.4 Final Notes

Time is the dimension through which human beings make sense of memories and experience. This is especially true of narrative, which is categorized by a sense of sequentiality. Without grounding in time, a model cannot understand or reason about narrative texts; this dissertation demonstrates that time can be linked directly to the underlying purpose and motivation of a narrative. It argues for an approach that is human-centered, and integrates elements of ground-truth through a better understanding of how reality relates to text.

The three perspectives used within this dissertation draw heavily on this human experience. This cross-perspective approach to time incorporates narrative knowledge and focuses on an example within the clinical text sphere of rich, diverse narrative. The corpus presented in this text and subsequent works center the human experience at every step. The process behind the generation of human-created text is integral to understanding the text's function, and while there is work yet to be done in translating this information to a machine model, the analyses present in this work demonstrate that narrative encodes a distinct layer of temporal behavior in text. The framework, corpus, and experiments support a pipeline where narrative knowledge can illuminate nuances of textual temporality, surface deviations, and explain how texts are ordered along the axis of time.

Machine learning models which train on text without human understanding are limited–like humans in a cave, unaware of the world beyond. To step outside it, we must look for data, models, and methodologies that challenge our understanding of text. Here, this work seeks to do so with a deep examination of the complex interplay between intention and text, medicine and narrative, rhetoric and time.

10 Appendix

10.1 Appendix I: Annotation

This appendix section discusses additional work from Chapter 4.

10.1.1 Visualization Tools

This section discusses visualization tools from the NIH EpiBio time annotation project.

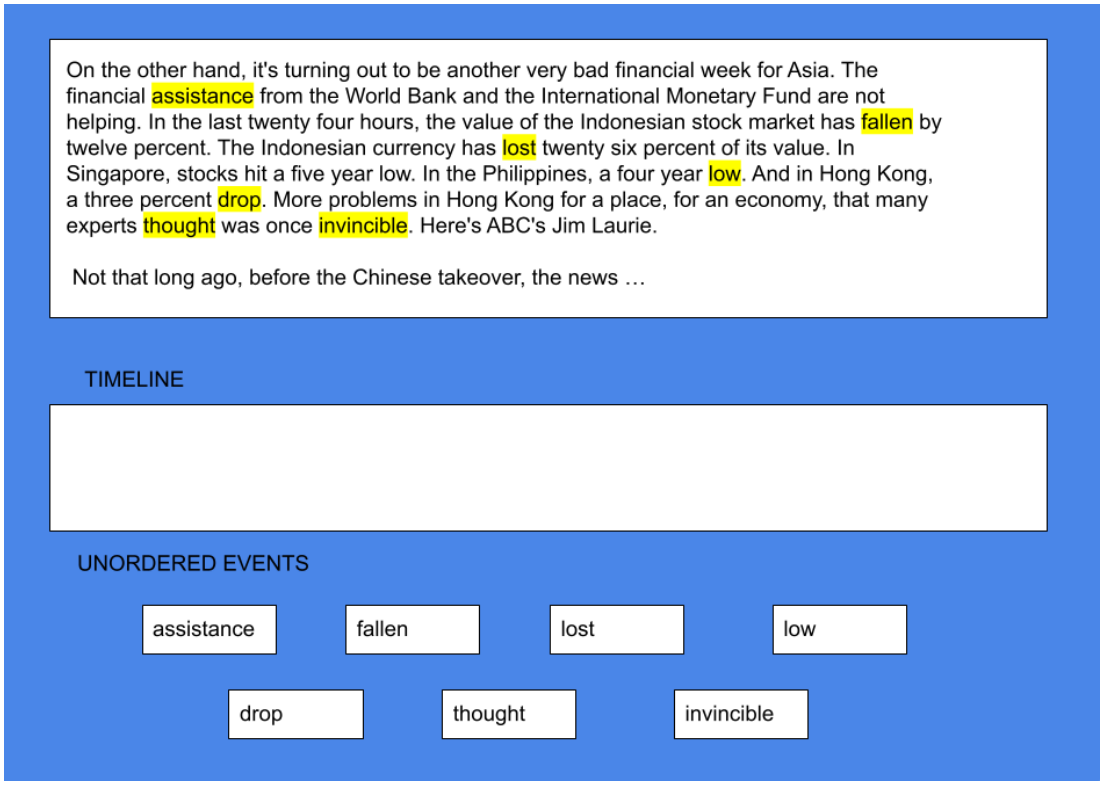

Figure 67: The layout of the Timeline Tool Version 1, containing a text display and incomplete timeline visualization.

The first iteration of the model (shown in Figure 67) was intended to bypass the $O(n^2)$ TEO/ETRE annotation cost by changing the label schema. Instead of asking annotators to label individual $Rel(E_i, E_j)$, an annotator could take individual events and place them directly on an incomplete timeline (reducing annotation bounds to $O(n)$).

As an example, suppose the incomplete timeline consisted of events $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4$, with E_5 unordered. If an annotator knew that $Rel(E_4, E_5) = before$, they could directly change the timeline to $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4 \rightarrow E_5$. Other types of annotation provide only partial information: if $Rel(E_3, E_5) = before$ and $Rel(E_4, E_5)$ is unknown, it is not possible to add E_5 to the timeline directly. However, that partial information might assist an annotator later in completing the timeline, so the tool allows for event-pair annotations in addition to the timeline annotations. These partial relations are tracked to extract entailments.

This type of temporal entailment logic (similar to what was used in TDDiscourse annotation in Section 4.2.4) was aimed to let annotators bypass 'challenging' pairs and quickly extract full annotations from partial information. In practice, annotators found events did not often fit easily into incomplete timelines, and the visualization chosen for partial annotations was unintuitive. The need for multiple types of annotation (event-level timeline annotation and pair-level ordering) required additional training and disrupted annotator workflows. Ultimately, this iteration did not improve time costs significantly compared to individual event-pair annotation.

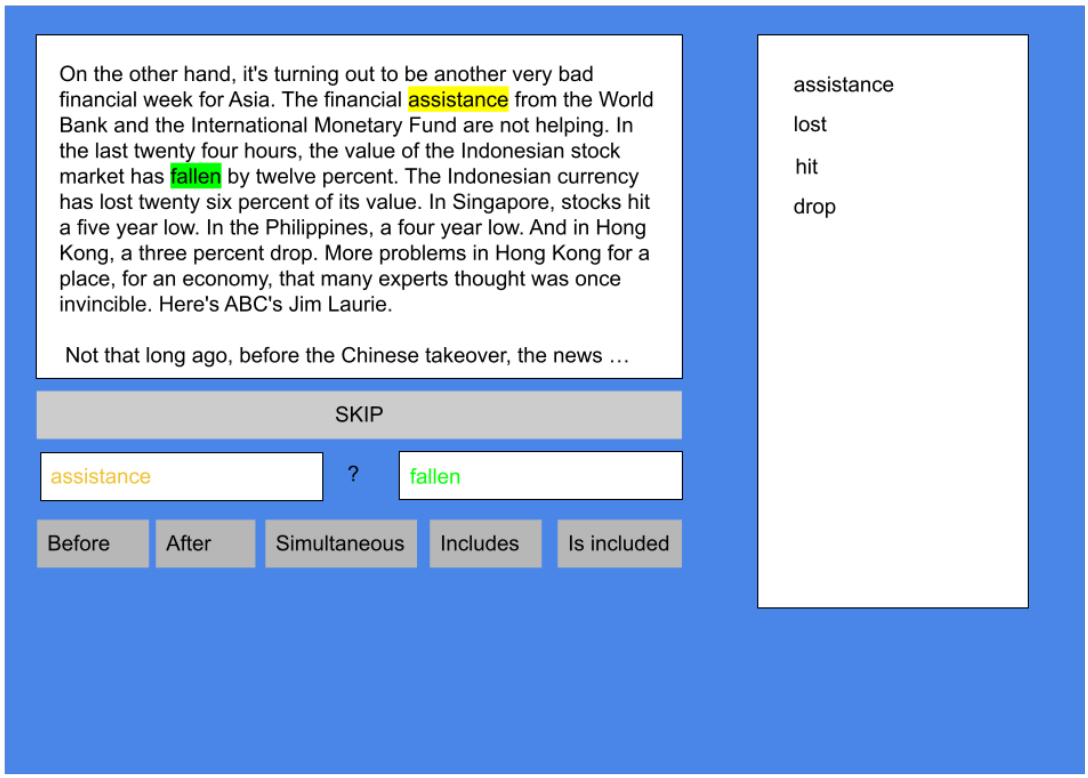

Figure 68: Timeline Tool Version 2. Note the pair-level annotation and incomplete timeline to the side.

The next iteration (shown in Figure 68) removed full event-level timeline annotation to focus on pair-level annotations. The tool's aim was to allow annotators to sort through event-pairs based on difficulty. Challenging pairs could be skipped with 'skip' button, with entailment information from simpler pairs providing information to help in labeling during later rounds of annotation. This is where the annotation workflow shown in Figure 39 from Section 4.3 is integrated. The tool presented human annotators with temporal entailments which followed from existing annotations. Annotators could confirm these entailments to add them to annotation output. This reduces the number of manual annotation required from annotators, though it is still upper bounded by $O(n^2)$.

In this iteration, the timeline is rendered vertically alongside the document text and pair information. The simple visualization had clear limitations for displaying a partial timeline.

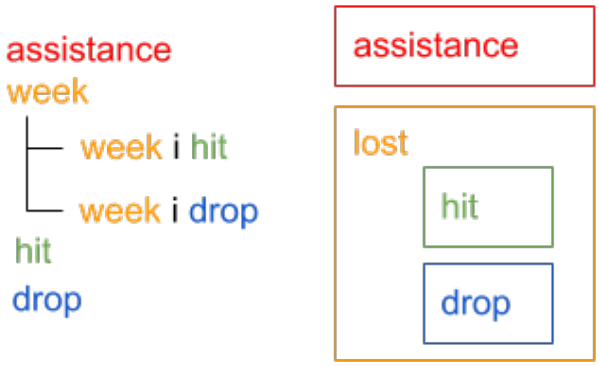

Figure 69: View of incomplete timeline (left), where annotators could list all existing pairs for a given event, and completed timeline (right), where before/after can be visualized as well as cases where events include one another.

It was difficult to render the complexity of partial relations between events; in order to render them within a linear display, the second iteration of the timeline tool was forced to choose one arbitrary ordering possible based on the information provided. Individual pair relations could be shown per event by selecting one event from the partial timeline (see the left-hand side of Figure 69 for an example of the interface). More detail was available to annotators once the timeline had been completed fully (see the right-hand display of the same figure). This allowed annotators to make final edits and confirm the timeline before annotations were saved and exported.

In practice, this tool was limited in how it could assist annotators. It was noted to be help annotators locate the context for event-pairs and reduce cognitive load, while the entailment process saved some effort across annotation rounds. However, annotators did not use the self-sorting workflow despite encouragement, the incomplete timeline display was actively unintuitive to the annotation team, and adjudication was still needed to produce Inter-Annotator Agreement that reached the level of 'fine'.

10.1.2 Timelines as Graphs

Basic TEO/ETRE Entailment:

A graph represents the simplest and most comprehensive data structure to capture relational data between nodes. In the case of a timeline, relations can be summed up as asymmetric, unidirectional edges from past to future: a **directed acyclic graph** (or DAG). This element of the work returns to logical properties across temporal relations formalized in TDDiscourse work (Naik et al., 2019): in addition to *logical entailments* between event-pairs, there are also *logical contradictions*, where event-pair annotations contradict a known entailment. Here, the salient contradiction is as follows:

- If $Rel(A, B) = before$ and $Rel(B, C) = before$, then $Rel(A, C) = after$ contradicts the known entailment.

These entailments capture physical properties of time; for any text representing real events, a contradiction of temporal ordering cannot be true. The contradiction type above reflects the most simple version of a graphical *cycle* that can be built using individual event-pair relations.

Any cycle will violate the necessary properties of a real timeline.

This effort define the nodes of a timeline DAG as representing one of the following:

- A single event on the timeline ($Node_X$).
- A node with children, representing a set of events (or start/end points) which occur simultaneously to one another ($SimNode$).
- The start or end point of a single event on the timeline ($Node_{X-start}$ or $Node_{X-end}$).

The flexibility of node representation in this framework allows the graph to use a single type of directional edges: a directional edge $Edge(source, target)$ indicates that $Node_{source}$ occurs before $Node_{target}$. Note that edges encode temporal direction and not distance. It is possible to represent any of the original 5 temporal labels using the following DAG logic:

1. $Rel(E_A, E_B) = before \rightarrow Edge(E_A, E_B)$
2. $Rel(E_A, E_B) = after \rightarrow Edge(E_B, E_A)$
3. $Rel(E_A, E_B) = simultaneous \rightarrow$ create $SimNode$ with children (E_A, E_B). Instructions for $SimNode$ creation are found below.
4. $Rel(E_A, E_B) = includes$:
   - Split $Node_A$ into $Node_{A-start}$ and $Node_{A-end}$. Instructions for node-splitting are found below.
   - Insert edges $Edge(Node_{A-start}, Node_B)$ and $Edge(Node_B, Node_{A-end})$ to capture the relation between E_A and E_B.

   A limitation of the TEO/ETRE 5-label annotation schema is that it is possible for $Rel(E_{B-end}, E_{A-end})$ to be $before$, $simultaneous$, or $after$.
5. $Rel(E_A, E_B) = isincluded$:
   - Split $Node_B$ into $Node_{B-start}$ and $Node_{B-end}$.
   - Insert edges $Edge(Node_{B-start}, Node_A)$ and $Edge(Node_A, Node_{B-end})$ to capture the relation between E_A and E_B.

   The same limitations are for $includes$ apply.

Complex Node Types:

The more complex node types require additional steps after creation to integrate them into an ongoing DAG.

When creating a $SimNode$ with children (E_A, E_B), the following is done:

- If E_A or E_B already belong to a $SimNode$, the children of the prior $SimNode$ become children of the new.
- For all existing edges of type $Edge(Node_X, Node_{A/B})$ (or in the case that E_A or E_B belong to a $SimNode$, the edges linking other nodes to the $SimNode$), edge becomes:

  $Edge(Node_X, SimNode)$
- For all existing edges of type $Edge(Node_{A/B}, Node_Y)$, edge becomes:

  $Edge(SimNode, Node_Y)$

Whenever a node (ex. $Node_A$) is split, the graph is automatically altered as follows:

- Create edge: $Edge(Node_{A-start}, Node_{A-end})$.
- For all existing edges of type $Edge(Node_X, Node_A)$, edge becomes:

  $Edge(Node_X, Node_{A-start})$
- For all existing edges of type $Edge(Node_A, Node_Y)$, edge becomes:

  $Edge(Node_{A-end}, Node_Y)$

In this way, the tool can can represent even complex timelines. Take Figure 70, for example, as a timeline with multiple branching paths:

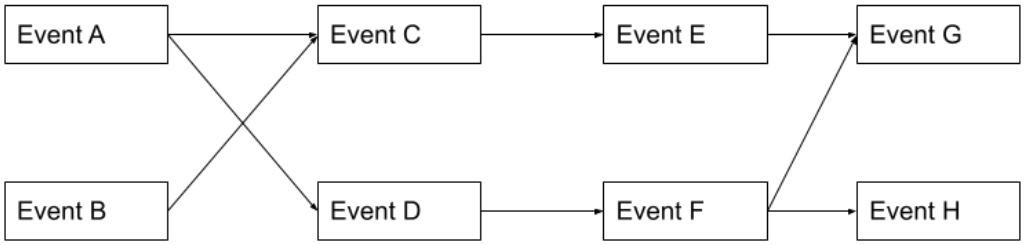

Figure 70: A complex timeline example, captured cleanly through DAG structure.

This timeline is difficult to render from a purely linear perspective. But a DAG captures all edges without loss of data.

Algorithmic Function:

The use of a DAG timeline allows this work to algorithmically implement the following functions which are useful and necessary to a timeline annotation tool:

1. Automatic annotation of logical entailments.
2. Identification of a completed timeline.
3. Capture of contradictory event-pair relation inputs.

Both prior iterations of the timeline collapsed the total number of manual event-pair annotations required for a comprehensive timeline. The pure timeline approach (iteration 1) did so using event-level annotations, and the event-pair approach (iteration 2) tracked specific entailments across the document. The DAG approach (current iteration) can leverage the fact that edges encode only direction and not distance to track entailments algorithmically. The tool performs the following for a given DAG to identify and extract *all* entailment of event-pairs (Function 1):

1. If a directional path (of any size) exists in graph from $Node_A$ to $Node_B$, there is an *entailed relationship* $Rel(E_A, E_B) = before$.

2. If a directional path (of any size) exists in graph from $Node_B$ to $Node_A$, there is an entailed relationship $Rel(Node_A, Node_B) = after$.
3. If $Node_A$ and $Node_B$ are children of the same $SimNode$, there is an entailed relationship $Rel(E_A, E_B) = simultaneous$.
4. If a directional path (of any size) exists in graph from $Node_{A-start}$ to $Node_B/Node_{B-start}$, and a path from $Node_B/Node_{B-start}$ to $Node_{A-end}$, there is an entailed relationship $Rel(E_A, E_B) = includes$.
5. If a directional path (of any size) exists in graph from $Node_{B-start}$ to $Node_A/Node_{A-start}$, and a path from $Node_A/Node_{A-start}$ to $Node_{B-end}$, there is an entailed relationship $Rel(E_A, E_B) = isincluded$.

This list of entailed relations for a partial timeline can be extracted with a single traversal through the entire graph. For each new $Node_Y$ reached, this traversal can identify $Rel(E_X, E_Y)$ for all $Node_X$ in the list of edges leading to $Node_Y$. $Node_Y$ is then added to the path for edges leading from $Node_Y$. Note that paths to and from $SimNodes$ give relation information for all child $Nodes$.

This algorithmic operation has $O(N + E)$ time complexity for N = number of nodes and E = number of edges. This complexity can be further reduced using techniques discussed later, and the algorithm mitigates the time cost of overall annotation. Like Iteration 1 of the timeline tool, this method places events directly on the timeline with either complete or partial information about the event's ordering. Each label manually provided by the human annotator therefore has potential to provide information for up to T distinct event-pair relations, where T represents the number of events currently on the timeline. As annotations are added, T approaches the total number of nodes N.

To determine when a timeline is complete (Function 2), can be achieved alongside the traversal to complete Function 1. This is done using the following axiom:

> A timeline is **complete** when, for every unique pair $Pair(E_A, E_B)$ in the timeline, $Rel(E_A, E_B)$ can be extracted from the graph.

The tool compares its set of pairs for which $Rel(E_X, E_Y)$ entailed by the current timeline against the total set of pairs. A graph where these sets match is complete. If the graph remains incomplete, more manual annotation is needed.

Lastly, the tool can algorithmically detect contradictory inputs (Function 3) using a similar but distinct method of graph traversal. In this traversal method, the tool tracks nodes already visited at each step of the graph. If at any point, a node is reached that has *already been visited*, then the graph contains a cycle that must be corrected to maintain the acyclic property of the DAG. All *logical temporal contradictions* in annotation will introduce observable cycles. This allows the tool to alert annotations in real time upon the input of a contradictory event-pair label.

DAG Simplification

As discussed, DAG traversal (used to perform all three necessary functions of the timeline tool) has a time complexity of $O(N + E)$. Though this is an acceptable level of complexity for the work, larger values of N and E could have a noticeable impact on the tool's processing speed. Additionally, a visualized DAG follows a specific logic that is less intuitive for many annotators. For adjudicators to easily understand existing timeline information for even complex timelines, the tool must 1) remove redundant edges to simplify the existing DAG; and 2) improve the

DAG visualization so that it appears more linear. The intent in this step is to produce as close to a final, linear timeline as possible at each step of the annotation–one edge leading into and exiting each node. This requires certain modifications to the DAG data structure to ensure a minimal loss of existing timeline information.

When traversing through the graph, it is possible to track the number of paths which exist between any $(Node_X, Node_Y)$. (Call this number $Paths(Node_X, Node_Y)$.) If $Paths(Node_X, Node_Y) \geq 2$, then the tool can examine these paths for redundant edges. There are two scenarios for which this value is possible, both shown in Figure 71:

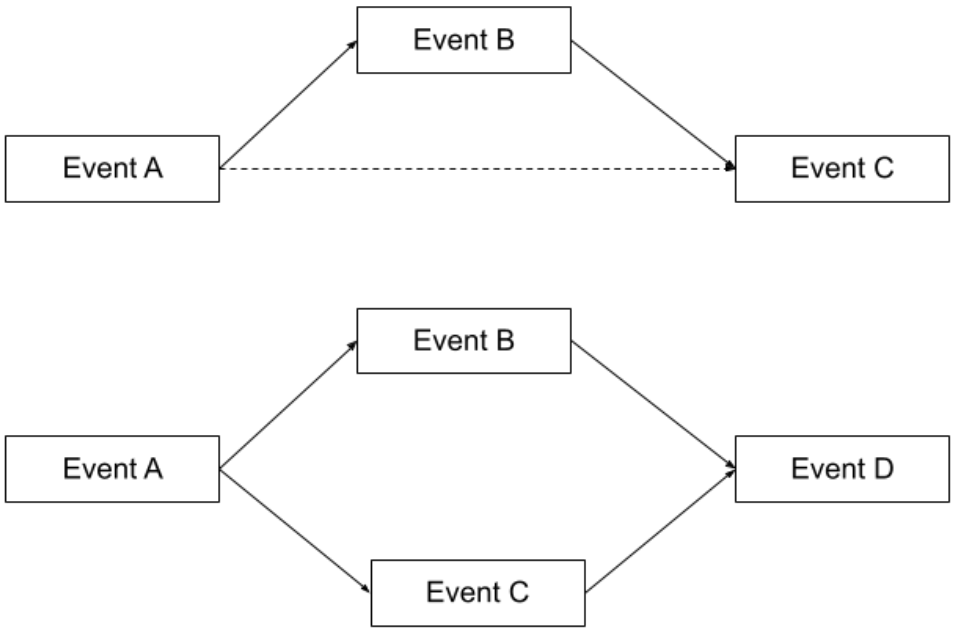

Figure 71: Two examples of DAGs with branching paths, one where a path is redundant (top) and one where each path provides unique data (bottom).

In each example, there are two paths which contain the same start and end nodes.

Graph 1 (top) has paths $Node_A \rightarrow Node_C$ and $Node_A \rightarrow Node_B \rightarrow Node_C$.

Graph 2 (bottom) has paths: $Node_A \rightarrow Node_B \rightarrow Node_D$ and $Node_A \rightarrow Node_C \rightarrow Node_D$.

In the first example, one of these paths is *redundant* because of how the tool's Function 1 calculates $Rel(E_A, E_C)$. The direct edge $Edge(Node_A, Node_C)$ can be removed with no information loss, as the new graph (see Figure 72) will still produce the same value of $Rel(E_A, E_C)$ after traversal.

The second example, however, contains no redundant edges. Though there are two paths from E_A to E_D, each contains unique information about events E_B and E_C. The graph therefore cannot be simplified through removing edges–a different method of reducing graph complexity is required to preserve the information already present.

To that end, the tool adds a second method of graph reduction: **non-simultaneous nodes**. A

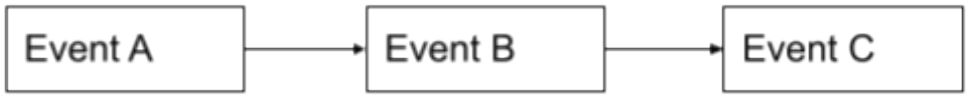

Figure 72: The direct edge from E_A to E_C has been removed. The DAG can be rendered linearly.

$NonSimNode$ contains a list of children such that $Rel(E_A, E_X) = before$ and $Rel(E_X, E_B) = before$ for all $Node_X$ in the list. The difference between a $NonSimNode$ and $SimNode$ is that the $NonSimNode$ is associated with its own DAG encoding relations for all children in the node. In this way, partial information can be stored about events in the non-simultaneous node while removing these edges from the main timeline until the sub-DAG is 'complete'.

Reducing this example produces the following (Figure 73):

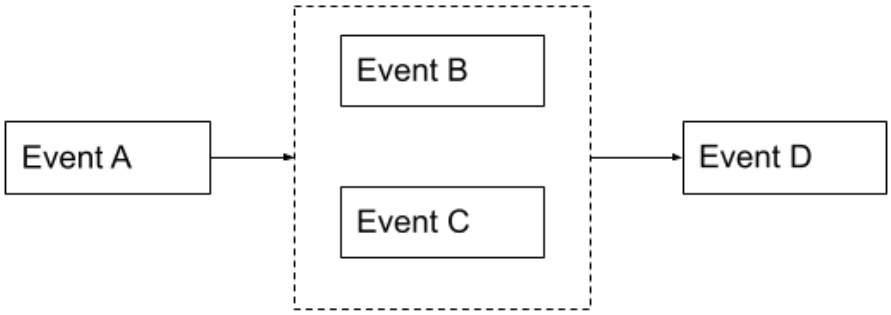

Figure 73: E_B and E_C have been moved to a $NonSimNode$ that preserves existing directional information.

There are limits to the utility of a $NonSimNode$[93], but they can capture many examples of even complex timelines. $NonSimNodes$ and their sub-DAGs can recurse infinitely, which makes them powerful tools in timeline reduction and visualization. Consider Figure 74, which uses nested $NonSimNodes$ and sub-DAGs to resolve the timeline.

The goal of $NonSimNodes$ is to reduce in-progress timelines to as close to a linear graph as possible. This speeds up the process of timeline visualization and ensures clearer understanding from annotators. The traversal step used for essential functions 1-3 can be easily adapted to account for sub-DAGs. This ensures that even partial information is fully processed during traversal.

[93]The example from Figure 70 is one edge case which can be rendered in a full DAG but not the simplified linear version.

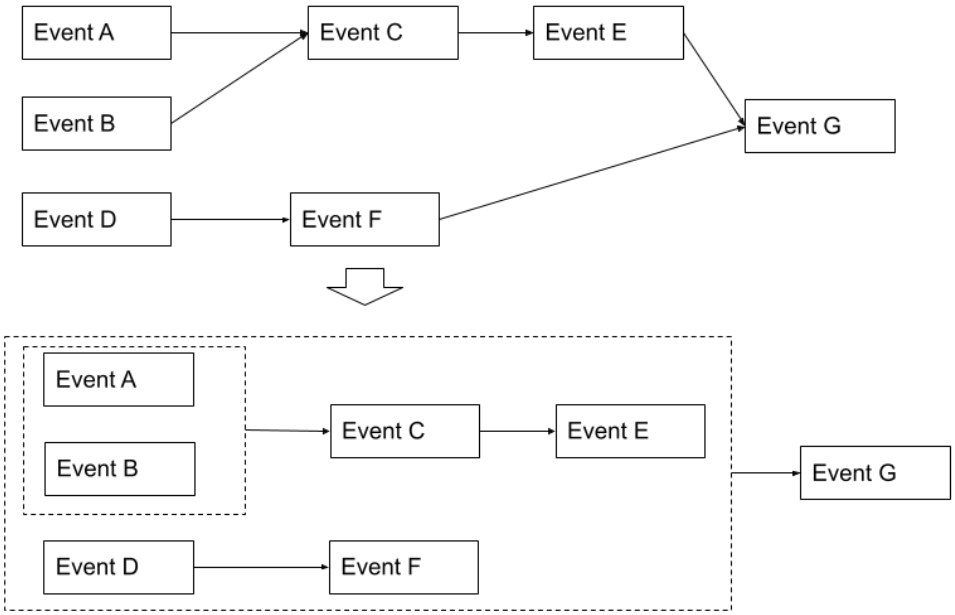

Figure 74: A more complex timeline resolved with $NonSimNodes$.

The user interface for Iteration 3 of the tool mostly matches the second, with the timeline visualization changed to a horizontal section beneath the text box (see Figure 75). These design decisions are justified based on feedback given in the last round. The ability to "skip" specific event-pairs remains and annotators are still encouraged to skip challenge cases, but the specific workflow of annotating in distinct rounds has been removed. Furthermore, there is no need to present entailed event-pair relations for annotator confirmation due to the DAG traversal process–all entailed relations are assumed to be correct.

The overall goals of Iteration 3 were: 1) to reduce the amount of manual annotations necessary for extraction through efficient extraction of entailed relation; 2) to visualize timelines in a way that synergizes with event-pair TEO/ETRE label schema, while building an informative and comprehensive reference timeline. The tool demonstrated in Figure 75 represents the best success for those purposes achieved within the NIH time annotation effort.

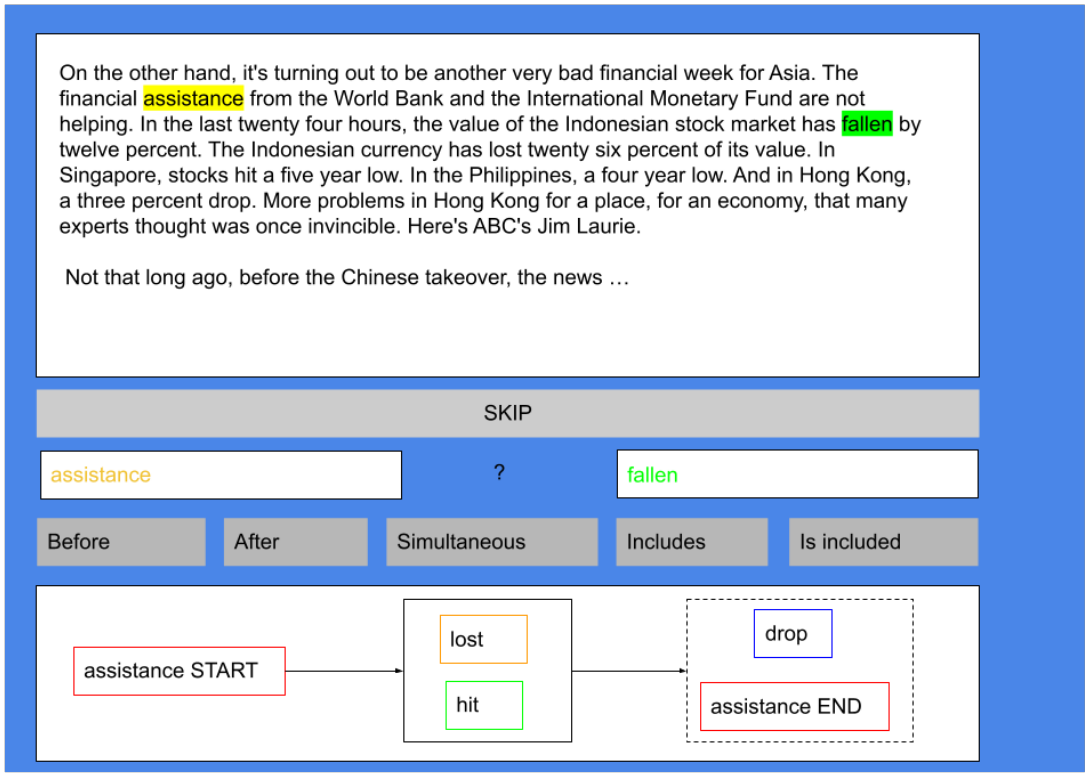

Figure 75: In the third timeline tool iteration, partial timelines display start and end points as well as simultaneous (enclosed in solid black box) and non-simultaneous (enclosed in dashed-line black box) node clusters.

10.2 Appendix II: IlNeS Survey

10.2.1 IRB Protocol

This section includes relevant sections of the protocol approved by Carnegie Mellon's Institutional Review Board:

Non-Exempt IRB Protocol For Non-Exempt (Expedited or Full Board review) research

Protocol Number: STUDY2024-00000013

Study Title: Study of Illness Narrative Structure Across Imagined Audiences

Principal Investigator (PI) Name: Luke (Marisa) Breitfeller

1. Study Scope. What is the purpose of the study (what is your research question/hypothesis) and how will the data collected be used to provide evidence for your hypothesis?

> "The purpose of this study is to compile a substantive corpus of individual patient illness narratives, with an eye towards 1) the imagined audience the narrative is constructed towards, and 2) the arrangement of temporal events within the constructed narrative. Our hypothesis is that the temporal ordering within the narrative will respond to the "imagined audience" of the work. To test this, we ask participants to share their illness narrative in response to distinct prompts and chronologically sort the events they include in the narrative. Illness narrative is a field of study which provides rich insight into human emotions and experiences when dealing with major life changes. The specific work we hope to do applies this insight to the general question of how human beings communicate the past in narrative text.
>
> The data will be used in addition to train models for temporal event-pair ordering on these provided narratives."

For each activity and subject population, clearly and completely describe the research procedures that will be conducted. In the SPARCS system, upload any questionnaires, surveys, tools, device manuals, pictures, links to videos, Manual of Procedures, test protocols, etc. that will be used to collect data or to direct the conduct of the study. Anything that will be seen by the subjects of the research or used to conduct the research must be submitted to the IRB for approval.

> "Participants will sign an electronic consent form when they participate in the study (multiple times, if applicable). This form (or forms) will be kept separate from the study records and answers not linked to the participant's name.
>
> To better accommodate participant's various schedules and disability, the study will be performed online, using a Zoom meeting to connect to the participants. In the first 20 minutes, we explain the protocol and allow participant to ask questions while signing the e-consent form. Protocol and e-consent form will be emailed to participant and displayed via screen-share. Video call will be used, but no recording will be made of participant audio or video. This will limit the reach of our study some, as it may exclude potential participants who do not have internet access or some type of computer. However, we note that we are specifically seeking participation from individuals with chronic or significant illness and disability, and as such it may not be possible for some in our potential participant base to respond to a survey in person.

> Participants may complete the survey in a number of ways. They may receive the study form via email to fill out on their own machine, and upload the completed form to this Box folder during the session. At no point in the Zoom call will their study responses be recorded through the Zoom interface. Participants may be granted write-only access to a sub-folder within the study's password-protected Box storage. This sub-folder will contain only their own study data, and a new sub-folder will be created for each new participant. The study researcher will have access to all participant sub-folders, but will use the sub-folders only to organize the raw data before additional anonymization and transfer to the final study dataset. Finally, participants may be given a link to a survey hosted on Qualtrics and distributed through the Prolific recruitment platform. Survey answers provided in this method will be hosted on Qualtrics before being transferred to the permanent Box folder and deleted from Qualtrics."

For each activity and subject population, indicate the location(s) where the research will be conducted. Specify whether the subject will be engaged in person, remotely via the internet, etc.:

> "Participants will be engaged remotely via the internet."

For each activity and subject population, describe the time required of the subject (time for each study visit AND overall time commitment for the study):

> "The discussion of protocol and consent will take place in the first 20 minutes of the session, and survey-answering in the remaining 40, for a total of 1 hour per participant. There is only one session for the study per participant."

Will questionnaires or surveys be used? *Yes.*

2. Participation Information

What is the age range of subjects in the proposed study? *18 years of age and above.*

How many subjects/records/specimens are needed for the study? Cannot be "unlimited".

> "We require > 35 participants per survey sub-section (150 total)[94]. Because there is only one session per the study, subjects are enrolled and complete the study within the same session and no withdrawals are expected."

How was the requested enrollment number determined? Provide power analysis or other justification for the number requested.

> "We are seeking a significant enough size n to make conclusions about overall patterns per sub-section of the survey."

Please list all inclusion and exclusion criteria for your selection of eligible subjects (include any age ranges, locations, abilities, experiences, or other qualifications that either include or exclude subjects from participating):

"Inclusion Criteria:

- Eighteen years or older
- Experiencing some illness, medical condition, or disability that has a significant impact on their life

[94] In practice, data collection stopped at 106.

- Comfortable writing in English
- Willing to share about their illness, medical condition, or disability in an anonymized format for this study
- United States citizens, US permanent residents, or in the US on a work or student visa
- Physically in the United States at time of survey

Exclusion Criteria:

- Medical condition directly impairs their ability to provide written survey examples in English
- Cannot access online survey tools"

What do you estimate the ratio of males to females to be? *The ratio of males to females is expected to be roughly one-to-one.*

Will any of the following vulnerable populations of subjects be involved in the proposed study? Select an answer for each population below. Please note that additional protections for any vulnerable populations will be required.

Pregnant women or human fetuses

> "Pregnant people would not be specifically excluded, but are not especially sought out for the study. A participant's involvement in the study is limited to the consenting adult taking a written survey which would not affect the fetus. Human fetuses are not an element of the study at all."

Neonates *No, all under 18 years are excluded.* Prisoners *No.* Children *No, all under 18 years are excluded.* Cognitively Impaired Adults *No.*

Students or Employees

> "Students would not be specifically sought out (recruitment materials will be placed in numerous off-campus locations), but would not be excluded provided they meet the other inclusion requirements. Students with direct connections to the PI and other researchers, along with CMU employees, would be asked not to participate."

Will the subjects be capable of understanding the nature of the study and the consent process? *Yes. We will select for participants with the cognitive capacity to understand the nature of the study and the consent process.*

If not, please explain: *N/A*

Will you target your research to enroll any specific demographics of the population, (e.g., race, ethnicity, sexual orientation, gender identity, religious group, etc.)? *Yes.*

Please explain:

> "We are specifically seeking survey responses from individuals dealing with some chronic or significant illness, disability, or other medical condition so that we can study how this specific population talks about their experiences."

Do you anticipate that your subjects will demonstrate an accurate representation of the population in the region where the study is being conducted? *Yes.*

If yes, please describe and estimate the percentage that will be from minority groups:

> "We are specifically seeking participants who have dealt with chronic illness or disability. Within that demographic, we expect participation from the Pittsburgh area to roughly correspond to Pittsburgh's population and for the online participants to reflect general demographics within the United States."

If no, please describe your study population and address why minority representation is not considered: *N/A*

Will subjects located outside of the United States be enrolled? *No.*

If yes, will specific countries be targeted for enrollment? *N/A*

3. Recruitment

Describe how subject recruitment will be performed:

> "We intend to reach out to patients through 1) physical flyers placed on CMU's campus, and local Pittsburgh community advertising bulletins (asking permission at each community site), 2) through advertising posts on online social servers and forums whose members often discuss illness and disability (again, as per permissions and rules of the associated forum/server), and 3) using survey site Profilic. We will also allow for snowball sampling."

Indicate how and by whom potential subjects are introduced to the study. Include any processes for screening for eligibility and conducting the informed consent discussion:

> "Participants will be given details through flyers, an online advertisement description, or from an acquaintance who had previously participated in the study. For the survey site Prolific, we will use the site's tools to screen specifically for our inclusion criteria and present the consent form with the survey."

Check all forms of recruitment that will be used for this study and upload all documents or language to be used in the Recruitment section of SPARCS:

- Flyers - State where will they be posted below. *Advertising bulletins in CMU's campus, and other local Pittsburgh area community bulletin boards.*
- Radio or TV *N/A*
- Email *N/A*
- Web-based - NOTE: If you are recruiting on mTurk, Qualtrics, Prolific or another similar online system, the title of the HIT/advertisement must include the SPARCS study number for identification of the study. *Through advertising posts on online social servers and forums whose members often discuss illness and disability (as per permissions and rules of the associated forum/server).*
- Subject Pool - State which one below. *N/A*
- Other - Describe below: *We will use survey site Prolific to allow interested participants on the site to engage with our survey. We will allow for participants to pass along recruitment information to other participants for snowball sampling.*

Will subjects undergo screening for eligibility prior to their participation? If yes, please note that you must request a Waiver of Documentation of Consent for screening purposes by completing the appropriate section within this form. *No.*

4. Consent

Do you plan to use consent forms? This includes any consenting language (written or verbal) provided to subjects prior to participation. *Yes.*

If yes, describe the process of how consent will be obtained, and by whom. Please include that consent will be obtained prior to any research procedures or data use/collection.

> "Participants will sign an electronic consent form when they participate in the study but before the survey begins. This form (or forms) will be kept separate from the study records and answers not linked to the participant's name.
>
> To better accommodate participant's various schedules and disability, the study will be performed online (for one-on-one survey participants), using a Zoom meeting to connect to the participants. In the first 20 minutes, we explain the protocol and allow participant to ask questions while signing the e-consent form. Protocol and e-consent form will be emailed to participant and displayed via screen-share. Video call will be used, but no recording will be made of participant audio or video. This will limit the reach of our study some, as it may exclude potential participants who do not have internet access or some type of computer. However, we note that we are specifically seeking participation from individuals with chronic or significant illness and disability, and as such it may not be possible for some in our potential participant base to respond to a survey in person.
>
> Survey participants recruited and performing the study through Prolific will be given the consent form alongside their survey, and will not have the one-on-one format with which to ask questions or request clarification. This is a limitation of asynchronous surveys."

Will the consent form be presented on paper? *No.*

Will the consent form be presented online? *Yes.*

Will the consent be presented verbally to subjects via a script? If yes, please complete the Waiver of Documentation of Consent section below. *No.*

Are you requesting to use a consent form that is different from the CMU template consent? *No.*

Are you requesting a Waiver of Informed Consent or an Alteration of Informed Consent? *No.*

Are you requesting a Waiver of Documentation of Informed Consent? *No.*

Will this study involve minors (children) as subjects? *No.*

5. Risks and Benefits

There must be sufficient benefit to conducting the research to outweigh any potential risks to participating subjects. Please be sure to list any potential direct or indirect benefits to subjects OR to the scientific community from the knowledge that will be gained by conducting the research and any potential risks to the subjects or others.

Will subjects receive a DIRECT BENEFIT from the study? Compensation for participation and experience with research-related technology or topics are not considered to be benefits. In many cases, subjects do not experience a direct benefit, but if they may, please list it here. *No.*

Indicate the expected INDIRECT BENEFITS to subjects, future individuals or groups, OR to the scientific community from the knowledge that will be gained. Please note that the research MUST have sufficient benefits to outweigh any risks.

> "A better understanding of the way patients with illness speak about their experiences, and in particular how the expected audience influences the temporal organization of the narrative."

Indicate all of the POTENTIAL RISKS to subjects. If any identifiable information from subjects will be accessed, used, recorded, or collected, please include a risk of Breach of Confidentiality:

> "This study focuses on a participant's individual experiences with chronic illness, terminal illness, or disability. These topics may be distressing to a participant to recall. Additionally, while researchers will not associate survey data with a participant's real name and will scrub specific identifying information from participant's answers, it is always theoretically possible that a participant's identity could be uncovered from a detailed personal story. Given the potentially sensitive nature of illness and disability, patients should understand that the risk, while small, is still present should they participate in this survey."

Indicate how each potential risk listed above will be managed and/or minimized:

> "To mitigate the risks, participants maintain full freedom over what is and is not included within their personal story, and we have no requirements that any particular details be featured. We inform participants they should not put any information in their story that they would not feel comfortable with us retaining in our study data.
>
> Participants will be given numbers of therapeutic hotlines in their consent form before the survey, with a note that such resources may be helpful if revisiting the memories asked about for the study have a negative impact on participants' health.
>
> In addition to allowing participants to withhold any details from their narrative, we take the following steps to preserve anonymity:
>
> 1. Study answers will be linked only to an anonymous identification number, not the participant's name.
> 2. In the written narrative, we encourage participants to avoid sharing overly-specific identifiable details, like names, dates, location names.
> 3. In transferring written records from initial collection to our permanent study database, any remaining details you leave in that we believe could be used to identify participants (ex. names, dates, location names) will be redacted from the stored records.
> 4. Any written records containing potentially-identifiable details will be redacted and destroyed after transferring records to the study database.
> 5. Records will be stored permanently in a password-protected Box folder accessible only to the research team."

Indicate the level of possible risk you believe the research will pose to human subjects (e.g., physical, psychological, legal, social, reputational, financial, etc.). *Minimal Risk.*

6. Deception or Incomplete Disclosure

Will deception or incomplete disclosure be used? *No.*

7. Compensation

Are subjects to be compensated for their study participation in any way? Compensation includes cash, checks, gift cards, course credit, parking, food/snacks, physical gifts, or chances in random drawings. *Yes.*

If yes, what is the value of compensation: *$15 per participant.*

If yes, what is the type of compensation (e.g., gift card, cash): If using a gift card, specify the retailer/type (e.g., Amazon, Visa, Target, etc.). Please also indicate how the subjects will receive their compensation (e.g., email, mail, in person, online platform such as Prolific) *One-on-one participants will receive Visa gift card via email, participants on Prolific paid through platform.*

Will subjects receive any non-monetary compensation (e.g., parking validation, snacks, chances in a random drawing)? *No.*

Are there any costs to subjects? Please include all potential costs related directly to their participation such as data charges for use of their phones, software licenses, membership fees, etc. These are costs that subjects only incur due to their participation in the research. *Yes.*

If yes, please describe:

> "It is possible a participant who can only access the survey through their phone, and who must pay data charges to access the internet from their phone, would incur charges. Potential participants will be told of the requirements for the study beforehand and we would not push participants to use this method if other, cheaper options are available to them."

Will you compensate subjects for injury resulting from participation? *This is an online survey. No.*

Will subjects who are students be offered class credit? *No.*

8. Data Security and Confidentiality

INITIAL identifiability state of research information – Select the type of information to be accessed/collected/obtained/used for this research. Please note that if you will ever have access to any information about subjects that can identify them (even if this information isn't being analyzed as "data"), your research is identifiable. *Identifiable.*

If identifiable, check each identifier accessed/collected/obtained/used: *Name (including initials), Email address.*

Describe the information being collected and/or used for the research. Provide a list of all data elements to be included in the research. You may provide a list here OR upload a spreadsheet or list of all data points into the SPARCS application in the Local Site Documents page.

> "Participants will reach out to the research team via email, and when obtaining consent, participants will have to provide their legal name. Consent forms including names will be available to the research team, but will not be linked to the participant's survey in the data set. No future work with the data set will allow access to the consent form. In the survey itself, participants will be discouraged from including identifiable data of the types listed above, including also names of acquaintances, places, and exact medical test results. If found in the survey results despite this warning, such identifiable data will be permanently redacted from the answers."

Will the research use existing data sets/recordings/specimens? *No.*

Describe your procedure for coding your data (encoding), if applicable:

> "We will obtain participant consent forms, which will require the participant's legal name. These forms will be stored separately from the survey data, and no link provided between the consent forms and the surveys themselves. Research team will review answers given in the survey itself to ensure no identifiable information is kept. Because the survey consists of freeform text, all text around the identifiable information will be preserved (ex. a survey answer might include 'I went with my friend Donna to UPMC Montefiore'[95], which would be encoded in our data as 'I went with my friend [REDACTED] to [REDACTED]'). Previous survey documents containing unredacted data will be permanently deleted."

Will AUDIO OR VIDEO RECORDINGS be made? *No.*

Did you obtain a Certificate of Confidentiality (CoC) from NIH? Note that research funded by NIH is automatically provided a CoC by NIH. Principal Investigators of non-NIH funded research may request a CoC from NIH, if desired. *No.*

In addition to the individuals listed on the study personnel page, who will have access to research data (e.g. surveys, questionnaires, recordings, interview records, etc.)? Include a comprehensive list and indicate if information may be shared outside the research team and/or CMU (including with collaborators, vendors, sponsors, etc.). Include what data each party will have access to and how the data will be transmitted/shared with them.

> "We intend to release the fully de-identified survey results after the project for work by future research teams. In this case, other research teams would only have access to the fully de-identified surveys, and not the set of consent forms which list our participant pool, nor will they have access to the emails originally used to contact the study."

Describe how you will protect subject confidentiality and secure research records (e.g. password protected, encrypted, etc.). Include location of where the data will be stored. If the PI should leave the university indicate your plan for the storage of research information and who will be responsible for oversight.

> "We will secure research records in a password-protected CMU Box account accessible only by our research team. Collection locations (like Qualtrics) will be regularly purged of research data. Should the PI leave the university, access and oversight will be left with the academic advisor for the project, Professor Carolyn Rose."

Describe your process for overseeing your study. Include a description regarding monitoring of data (to ensure that study goals are met and adherence to the IRB-approved protocol is maintained). Examples: Review of lab notebooks, frequency of meetings to review data, who will be present at the meetings, how recruitment and retention will be monitored, etc.:

> "I as the PI will be meeting with my academic advisor weekly and can discuss study goals and adherence to IRB protocol there. We can discuss recruitment for the study (given each survey requires only one session, retention is not an issue) and if any concerns have come up from participants. We will ensure all data remains secure in the designated Box folder."

[95] A synthetic example of identifiable data.

Describe your process for ensuring that adverse events, unanticipated problems, and subject complaints are reported to the IRB Office in a timely manner. Please note, all reportable events are required to be reported to the IRB via a Reportable New Information (RNI) form in SPARCS within 3 days.

> "The consent form given to participants provides options to contact CMU directly for complaints about the study process, and we will also ensure participants can reach out to the PI through the original study email contact. Any reportable participant complaints or other adverse events will be communicated to the PI's academic advisor and a report made to SPARCS within 3 days."

Please describe the intended FINAL IDENTIFIABILITY STATE of all information collected/ obtained/used/generated for this study (e.g., Will you initially be collecting identifiable information for this research, but during the course of the study you will de-identify it?). The FINAL state of data at the end of the study will be: *De-identified*

9. External Collaborators (not CMU-Affiliated)

Is this research to be done in collaboration with any institutions, individuals, or organizations that are not affiliated with CMU? *No.*

Is there or will there be IRB approval from another IRB for this study? *No.*

10.2.2 Timeline Leaf/Step Examples

This section of the appendix shows examples of the "timeline leaf" paradigm from within the IlNeS corpus. For each timeline, the "step graph" is also provided to show how this methodology may surface inter-leaf interactions and other deviations.

The first example of the leaf paradigm (shown in Figure 76) uses Timeline 6282 from IlNeS. Timeline 6282 can be split into three leaves, each of which is chronological with respect to itself. These sections of the timeline can be interpreted as covering three distinct semantic topics: 1) "before symptoms start", 2) "time during symptoms", and 3) recovery. Each leaf timeline is chronological with respect to itself; all deviations in full timeline are due to inter-leaf interactions.

In the step graph (Figure 77), most deviation appears in the inter-leaf interaction zone (marked in red). Significant deviation exists at positions 1-4, 6, and 13-14, consistent with the steps in the timeline which switch between leaves. One case of 'local' deviation is surfaced at position 9. In this case, it does not reflect a true local deviation (ex. where two adjacent events are presented in reverse order), but movement from one timeline leaf to another that happens to be small ("socially impaired" and "pretend to be sick" are near one another in the full timeline, but are best conceptualized as belonging to distinct timeline leaves).

Figures 78 and 79 represent Timeline 7542. This timeline contains two leaves: Leaf 1 features a *timeline inversion* and minor ordering deviations, while Leaf 2 contains local deviations. These leaves can be interpreted as covering topics: 1) past medical history and 2) current symptoms. Timeline 7542 is an example of a *discrete* multi-part timeline compared to prior examples which are *disjoint*–that is, the timeline does not jump between leaves. Further, 7542 presents each timeline leaf in the same order that they occurred in reality; the only difference between leaves is their underlying timeline attribute (inverted versus chronological).

The step graph of Timeline 7542 therefore showcases something vital about this methodology: common timeline phenomena a steps-graph can surface, and how they *interact*. Each segment

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"took Lexapro"	"broke up"	"quit"	"got PSSD"	"emotionally numb
Textual:	"got PSSD"	"took Lexapro"	"emotionally numb	"broke up"	"quit"
	Position 6	**Position 7**	**Position 8**	**Position 9**	**Position 10**
Chronological:	"woke up"	"relentless craving	"socially impaired"	"socially withdrawn"	"pretend to be sick"
Textual:	"woke up"	"relentless craving	"socially impaired"	"pretend to be sick"	"pleasure began to return"
	Position 11	**Position 12**	**Position 13**	**Position 14**	**Position 15**
Chronological:	"pleasure began to return"	"panic attacks"	"dropped out"	"was arrested"	"recovered"
Textual:	"panic attacks"	"dropped out"	"socially withdrawn"	"was arrested"	"recovered"

Leaf 1	**Position 1**	**Position 2**	**Position 3**			
Chronological:	"took Lexapro"	"broke up"	"quit"			
Textual:	"took Lexapro"	"broke up"	"quit"			
Leaf 2	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**Position 6**
Chronological:	"got PSSD"	"emotionally numb	"woke up"	"relentless craving	"socially impaired"	"socially withdrawn"
Textual:	"got PSSD"	"emotionally numb	"woke up"	"relentless craving	"socially impaired"	"socially withdrawn"
Leaf 3	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**Position 6**
Chronological:	"pretend to be sick"	"pleasure began to return"	"panic attacks"	"dropped out"	"was arrested"	"recovered"
Textual:	"pretend to be sick"	"pleasure began to return"	"panic attacks"	"dropped out"	"was arrested"	"recovered"

Figure 76: Timeline 6282 with leaves labeled.

of a timeline has two key attributes, **direction** and **deviation**. Though their interactions may seem complex, they are in fact regular and mathematically predictable.

A timeline segment with chronological direction is represented by the values $x = 0, 1$ for all positions in sequence. (x will only equal 0 in the case that some set of events are *simultaneous* in real time, and only for the positions where the timeline moves between simultaneous events.) In an inverted segment of the timeline (where the segment has length m), the steps methodology will surface a sub-sequence (length $m+1$) of values $m, -1, -1, -1, ..., m$. The positive elements of this sub-sequence represent the significant jumps forward in time that must precede and follow the inverted sequence, and the consecutive steps of size -1 show steps backward in time.

Deviations are applied after these initial directional elements and add to the step values at each position. A simple swap between two adjacent events (when applied to a chronological subsequence) contributes values $1, -2, 1$ (where the start of the subsequence x is the chronological position of the first event in the swap). This can be seen in Figure 79 at Positions 13-15– consistent with events "take breaks" (Position 13 in chronological time) and "pass up" (Position 14) which are swapped in the text. A simple adjacent swap in an inverted sequence inverts this value contribution, providing the values $-1, 2, -1$. Therefore, the mathematical values of the full sequence **direction**, where the timeline is inverted from positions 1-4 and chronological for 5-15:

$$T_{dir} = 4, -1, -1, -1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$$

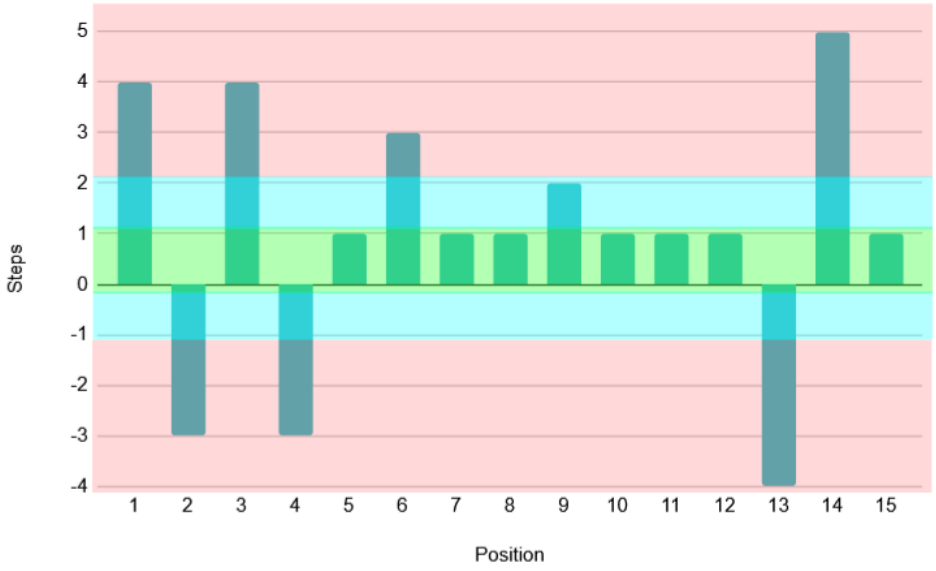

Figure 77: Steps visualization of Timeline 6282.

The step values contributed by each individual sequence deviation is:

$T_{dev3-5} \Rightarrow 0, 0, -1, 2, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0$

$T_{dev5-7} \Rightarrow 0, 0, 0, 0, 1, -2, 1, 0, 0, 0, 0, 0, 0, 0, 0$

$T_{dev13-15} \Rightarrow 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, -2, 1$

These values sum together even when deviations overlap. It produces the final timeline step-sequence of:

$Timeline \Rightarrow 4, -1, -2, 1, 4, -1, 2, 1, 1, 1, 1, 1, 2, -1, 2$

This matches what is shown in the step-graph. Therefore, this method presents an analysis which may immediately appear chaotic but actually encodes precise and regular operations explaining the edits made in the change from chronological to text.

Lastly, the example of Timeline 2768 (in Figures 80 and 81) shows an example of a highly-disordered timeline. Almost all steps in the step-graph surface a shift between leaves, indicating that leaves are likely small and require more granular topic categories to explain the final timeline. This type of timeline does not respond as well to the direction/deviation approach as the prior timeline; steps between timeline leaves are, on the whole, more chaotic and less regular than simple deviation types.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"military"	"pain"	"shin splints"	"arthritis"	"dull ache"
Textual:	"arthritis"	"shin splints"	"military"	"pain"	"walk"
	Position 6	**Position 7**	**Position 8**	**Position 9**	**Position 10**
Chronological:	"walk"	"sharp"	"swelling"	"arthritic pains"	"running errands"
Textual:	"dull ache"	"sharp"	"swelling"	"arthritic pains"	"running errands"
	Position 11	**Position 12**	**Position 13**	**Position 14**	**Position 15**
Chronological:	"going out"	"doing chores"	"take breaks"	"pass up"	"new doctor"
Textual:	"going out"	"doing chores"	"pass up"	"take breaks"	"new doctor"

Leaf 1	**Position 1**	**Position 2**	**Position 3**	**Position 4**	
Chronological:	"military"	"pain"	"shin splints"	"arthritis"	
Textual:	"arthritis"	"shin splints"	"military"	"pain"	
Leaf 2	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**
Chronological:	"dull ache"	"walk"	"sharp"	"swelling"	"arthritic pains"
Textual:	"walk"	"dull ache"	"sharp"	"swelling"	"arthritic pains"
Leaf 2	**Position 6**	**Position 7**	**Position 8**	**Position 9**	**Position 10**
Chronological:	"going out"	"doing chores"	"take breaks"	"pass up"	"new doctor"
Textual:	"going out"	"doing chores"	"pass up"	"take breaks"	"new doctor"

Figure 78: Timeline 7542 with leaves labeled.

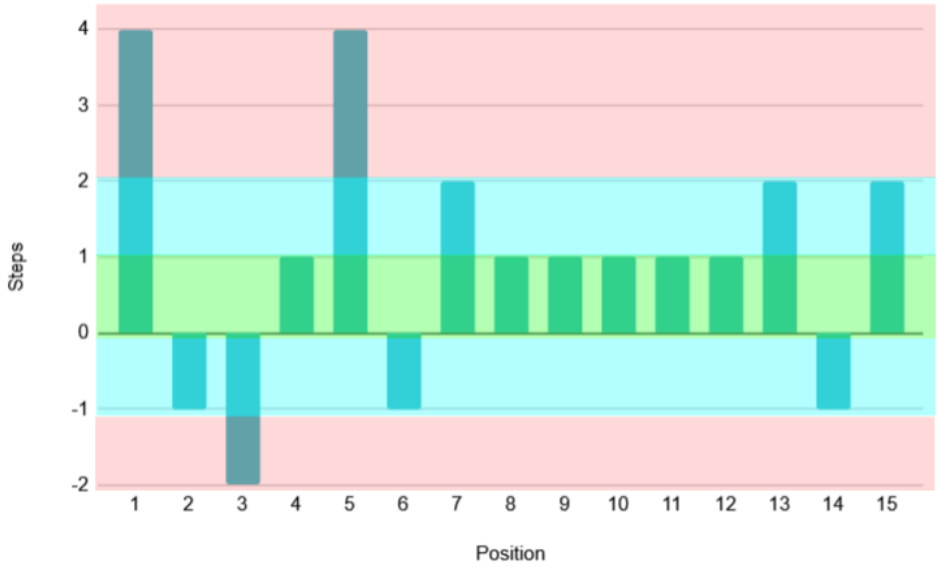

Figure 79: Steps visualization of Timeline 7542. Significant deviations can be found at positions 1, 3, and 5. The sequence of negative steps indicates an inverted portion of the timeline.

	Position 1	Position 2	Position 3	Position 4	Position 5
Chronological:	"grieve"	"lonely"	"support" (1)	"manage"	"doing more"
Textual:	"manage"	"enjoy the little moments"	"doing more"	"grieve"	"living a full life"
	Position 6	**Position 7**	**Position 8**		
Chronological:	"support" (2)	"enjoy the little moments"	"living a full life"		
Textual:	"support" (1)	"lonely"	"support" (2)		

Leaf 1	**Position 1**	**Position 2**	**Position 3**
Chronological:	"grieve"	"lonely"	"support" (1)
Textual:	"grieve"	"support" (1)	"lonely"
Leaf 2	**Position 1**	**Position 2**	**Position 3**
Chronological:	"manage"	"doing more"	"support" (2)
Textual:	"manage"	"doing more"	"support" (2)
Leaf 3	**Position 1**	**Position 2**	
Chronological:	"enjoy the little moments"	"living a full life"	
Textual:	"enjoy the little moments"	"living a full life"	

Figure 80: Leaf version of Timeline 2768. Semantically, these sections can be interpreted as negative emotion, symptom management, and positive recovery. Note that for highly-disordered timelines, contiguous leaves become very short.

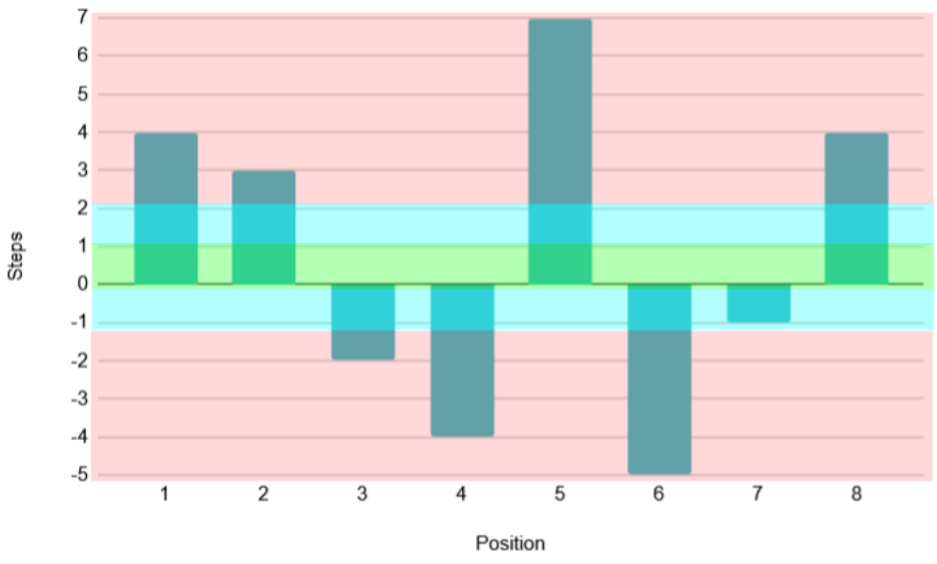

Figure 81: Steps visualization of Timeline 2768. A majority of steps in this timeline are in the "minor" or "significant" deviation zones.

10.3 Appendix III: IlNeS Analysis

This appendix demonstrates that the data collected in IlNeS represents a random sample of those living with illness or disability (excepting the inherent limitations enforced on recruitment)[96]. As a result, conclusions discussed in this chapter should generalize across that population. This appendix will discuss the IlNeS recruitment process, the spread of participants, and cases where data was limited by participant unwillingness to discuss certain personal or emotionally-charge topics, along with other logistical details.

10.3.1 Participant Behavior and Variety

As detailed in the protocol (see 10.2.1) the IlNeS work intentionally did not collect demographic data (specific age, gender, location, level of education, and so on) for our participant dataset. Though this data could provide rich and insightful information regarding how testimonials are shaped for each demographic, there are two reasons not to collect it.

1. The thesis of the IlNeS work is that intentionality affects text. Data must be collected to support that fundamental hypothesis before moving on to more specific experiments. If such a link exists, it is likely that trends in this behavior exist across demographics. But to prove the validity of the dissertation thesis, it is better to collect a range of data from multiple representative demographics to allow for generalizable conclusions.

2. This survey deals in highly private data. Personal health status is a class of data considered to be *personally sensitive*. A breach of that data could lead to discrimination and loss of status in everyday life, and is in general a violation of participant privacy. Re-identifying individuals from demographic data can be achieved with as few as three personal details and this risk of re-identification increases when details are highly specific (see Sweeney, 2000), like official diagnoses of uncommon illness. Therefore, IlNeS opted to lessen the risk of re-identification by taking no additional demographic data.

Represented Conditions:

The work's definition of "medical condition" is broad. IlNeS allowed participants to submit testimonials for any health condition, physical or mental, that could be considered chronic, long-term, or acute enough to have a significant impact on daily life. Though the survey did not require participants to be specific about the condition they have dealt with or its causal factors, many relayed that information regardless. In Section 5.5, these conditions are clustered into broad *illness types*, but this section provides a more comprehensive list. Note that this survey lists conditions which participants identify as 'disabilities'. The goal of the survey is not to pass judgment on what is and is not a disability, only to provide a forum for individuals to speak about their experiences. See Table 41 for a full accounting of reported conditions.

Condition	Source(s)	Fact.	Pers.	Emot.	Clin.
ADHD	Not specified	1	2	0	0
Allergy to cold	Environmental factors	0	0	0	1
Amputation	Injury	0	0	0	1
Angina	Environmental factors	0	1	0	0
Anxiety	Not specified	2	2	1	3
Apnea	Not specified	0	0	1	1

[96]Participants had to be 1) permanent residents or citizens of the United States, 2) fluent and comfortable in English. IlNeS results may only be valid within the culture of English-speaking United States residents, and other cultures may exhibit different trends within this space.

Arthritis	Environmental factors	0	1	0	2
Asthma	Congenital, environmental factors	0	0	0	1
Atrial fibrillation	Not specified	0	1	0	0
Autism	Not specified	0	0	1	0
Autoimmune disorder	Not specified	1	0	0	0
Blood clotting disorder	Not specified	0	0	0	1
Cancer (lymphoma and other)	Not specified	1	2	2	1
Celiac disease	Not specified	0	1	0	0
Chemical sensitivity	Environmental factors	1	0	0	0
Chronic abdominal condition	Not specified	0	1	0	0
Chronic anhedonia	Congenital	0	0	1	0
Chronic fatigue	Environmental factors	2	0	0	1
Chronic pain	Injury	3	0	0	2
Colon issues	Not specified	0	1	0	0
COPD	Environmental factors	0	1	0	0
Crohn's disease	Not specified	0	1	0	1
Depression	Congenital, environmental factors	7	1	3	4
Diabetes	Not specified	2	4	1	4
Dysautonomia	Congenital	0	0	0	1
Dyslexia	Not specified	2	0	0	0
Eating disorder	Not specified	0	0	0	1
Ehlers-Danlos	Congenital	0	0	0	1
Emphysema	Environmental factors	0	1	0	0
Endometriosis		0	1	0	0
Epilepsy	Not specified	1	0	0	0
Fibromyalgia	Not specified	0	0	1	0
Gut issues	Environmental factors	1	0	0	1
Heart disease	0	0	0	1	
High cholesterol	Not specified	0	0	0	1
Hypertension	Environmental factors, idiopathic	0	1	0	2
Hypothyroidism	Not specified	0	0	1	0
Inflammation	Environmental factors	1	0	0	0
Keratokonus	Not specified	0	0	1	0
Lobe infracts	Environmental factors	0	0	0	1
Long COVID	Environmental factors	1	0	0	0
Lung issues	Environmental factors	1	0	2	0
Mast cell disorder	Congenital	0	0	0	1
Migraines	Not specified	1	0	2	0
Mitral valve prolapse	Not specified	0	0	1	0
Myasthenia gravis	Not specified	0	0	1	0
Obesity	Not specified	0	0	1	1
Obsessive compulsive disorder	Not specified	1	0	0	1
PCOS	Not specified	1	0	0	0

Pineal cyst	Not specified	0	0	0	1
Postpartum depression	Not specified	0	0	0	1
Post-SSRI Sexual Dysfunction	Environmental factors	0	0	1	0
Psoriasis	Not specified	1	0	0	0
PTSD	Trauma	0	4	2	1
Pulmonary artery disease	Environmental factors	0	1	0	0
Rare genetic disorder	Congenital	0	1	0	0
Schizophrenia	Not specified	0	1	0	0
Scoliosis	Not specified	0	0	0	1
Stutter	Not specified	1	0	0	0
Traumatic brain injury	Injury	0	0	0	1
Uterine issues	Not specified	0	0	1	0
Vitamin D deficiency	Not specified	0	0	1	0
Not specified (physical)	Not specified	0	0	2	1
Not specified (mental)	Not specified	0	1	0	1
Not specified	Not specified	4	4	8	2

Table 41: Conditions used by IlNeS participants for testimonials.

Note that some conditions can be very rare, and the diagnosis or treatment plan alone could be re-identifying. As such, the research made redactions to the data in cases where a treatment plan seemed highly specific, and for cases where a medical condition was rare enough to risk participant privacy. Therefore, some conditions (like "rare genetic disorder"), are only referred to as such in the corpus itself and in subsequent analyses.

Analysis:

In 20 testimonies, participants chose to be vague with the nature of their condition. Two participants relayed that they specifically chose to omit details about their condition due to the "vague approach" advised in the instructions.

As discussed in Section 5.5, the work considered the possibility of correlations between the audience prompt given to the participant and the medical condition that is disclosed in the testimony. Though all prompts were assigned at random, it is common for individuals dealing with long-term illness to have comorbidities. Because the survey asked participants to choose one condition to focus on in discussion, the prompt may have influenced which condition in chosen in case of comorbidities. The analysis in that section did not find strong correlations between broad illness type and NarraType for testimonials which could be measured for temporal disorder[97], though it is possible some trends exist for specific conditions or within testimonials that lack an annotated timeline.

Omissions:

The survey data deals with sensitive and identifiable information. Therefore, it was essential to ensure participants had the means to mitigate potential discomfort and risk to their own person inherent to such disclosure. To that end, IlNeS allowed participants to choose which details to

[97]A testimonial must have a participant-annotated timeline of length $n \geq 2$ to be usable for analysis of disorder.

disclose, and to omit any which felt too distressing or personal to them. The instructions stated explicitly that participants were not required to inform the research team if they were making an omission, or what that omission might be. Though this approach limited data collection, it was in line with ethical standards common to narrative survey work.

Even with these instructions, some participants used the freeform response section to volunteer information about what they had chosen to omit, or that they felt they did not make omissions. This information is cited in Table 42.

Participant ID	Prompt Given	Feedback
0448	Clinician	"I avoided specifics about the condition itself since the prompt asked for a general approach."
1461	Persuasive	"I described my illness the way it progressed without providing too many details on the specifics of the condition"
1510	Factual	"I was very transparent and very open-minded to the answers I provided."
2111	Persuasive	"I didn't go into specific medical details, because I assumed the goal was to frame the broader consequences"
2276	Factual	"I remembered the teaching to not use identifying information, and instead of calling the condition by name, I explained that the symptoms are hard to predict, and that they impact my planning or ability to make commitments."
2492	Persuasive	"Was trying to recall all important information about the event without disclosing personal details [...]"
4399	Clinician	"I tried to give a full picture without going into too much detail."
5313	Factual	"I thought a long time about how much detail I wanted to actually go into and about how many specific details and events from my own story I wanted to put in."
6282	Emotional	"I did leave out some of the more graphic details about sexual health, though."
7310	Emotional	"[...] without getting too far into specific details [...]"
9460	Persuasive	"What I did not do I did not include internal step-by-step reasoning or raw deliberations about alternative phrasings. I did not invent any medical events, dates, or diagnostic detail that you didn't provide."

9551	Factual	"I tried to avoid bogging down the story with details I don't really care about anymore or would be long tangents (for example: bullying that took place through childhood up through high school)."

Table 42: Omissions reported by IlNeS participants.

Register and Vernacular:

Certain past works on formal surveys (Labov, 2013; Sneller et al., 2023) express concern that the interview style prevents researchers from collecting truly 'natural' language. Evidence in these past works points to an absence of the vernacular register in interview as a sign of inauthenticity among participant responses. Contrary to these concerns, IlNeS work found that in practice it is difficult to fully remove vernacular from survey responses. Participants drew from a wide range of registers and styles across the work (likely due to inherent individual differences), even when the register of the prompt did not change. Some participants used organized and polished language, with a clear understanding of formal argumentative structure. Others used vernacular language, writing in stream-of-thought, leaving in misspellings and grammatical errors.

Despite expectations, self-censorship is not an inevitability even in formal survey environments; many people simply will not stray from a vernacular style, no matter an interview scenario is presented.

10.3.2 Design Validation

This section shows the full collection of framing statements and feedback remarks which validate the choice in IlNeS design to prompt participants using proxy cues instead of directly soliciting NarraTypes, as discussed in Section 6.1.

Factual Scenario:

Framing statements and feedback notes which express some manner of intention from the Factual prompt are found in Tables 43 and 44. In some cases, these intention statements do not directly reference actions defined as containing clear factual intent (like "explaining" or "informing") but provide useful information about how participants interpreted the proxy prompt.

Participant ID	Framing statements
0775	"Hey, I have never really explained this to you before"
2276	"I would like to articulate a little of what I have been experiencing, since to the outside it may not make much sense."
5056	"I would like to share something about myself that not everybody knows about me. [..] It is not something I talk about a lot, but it is definitely a part of my background."
6611	"Hey, I don't think we've ever talked about this"

Table 43: Framing statements given in Factual narratives.

Participant ID	Feedback
0190	"I just wanted it to feel real, like something I might actually say if someone asked me"
1510	"I am trying my best to be very truthful to these conditions whiles I am hopefully having treatments for it."
2276	"I was considering the way I should explain my experiences to someone who had never heard of my condition. [...] In general, my thinking was connected to the idea of making the story relatable, respectful and personal without making it over-detailed and technical.
2609	"I tried to explain my depression as if I was explaining it to someone who has never had experience with it [...] I wanted to try to put things in perspective and simplify as best as I could for a maximum understanding of a complex issue."
4578	"I focused on conveying the personal experience of living with a condition that I wasn't initially aware of."
4677	"I was just thinking about my own personal experience with depression as it is a disability for me"
5056	"I wanted to convey that it was a difficult event and how it contributed to shaping me."
5313	"I wasn't really sure what I wanted to say or how I would even describe such a thing, especially to someone who doesn't even know what depression is."
6611	"I focused on being honest and clear about what it's like to live with a chronic illness. [...] My goal was to keep the tone real and relatable, showing that while it's challenging, I'm still adapting and learning to manage things."
6657	"I also thought about how best to express my painful experience so others can understand it. Most people didn't have to struggle with dyslexia so they don't understand what it felt and feels like."
8196	"I started by considering how best to explain dyslexia in a way that someone with no prior understanding could easily grasp. I focused on clarity, wanting to keep it simple while still being accurate [...] since it's a conversation with a friend, I aimed for a casual, relatable tone, something that would make my friend feel comfortable and open to asking questions if they were curious. [...] It wasn't just about describing dyslexia, but also about making it personal and relatable to someone outside the condition."
9401	"I thought about how I'd explain PCOS to someone who genuinely didn't know anything about it. I wanted to be honest and clear."
9508	"I wanted to explain my illness in a simple and straightforward way to someone who might not be familiar with it."

9551	"I tried to keep it straightforward and simple when explaining. [...] I tried to remain focused on the progression of the illness itself, rather than go through the potential causes."

Table 44: Framing statements given in Factual feedback.

Persuasive Scenario:

Persuasive NarraType framing statements and feedback notes are found in Tables 45 and 46.

Participant ID	Framing Statements
0848	"Thank you so much for considering allocating funds/resources into researching better treatment options for celiac disease."
1171	"To whom it may concern"
1464	"Hi, I write to you in times of desperation and just wanting to live a normal life. [...] I urge you politely and respectfully to allocate funds for better treatment and accommodations for individuals like me so that we may have a chance of living a somewhat normal and quality life."
1561	"I believe that you should consider allocating funds to researching my illness."
1933	"We need clearer diagnostic tools, smarter training for doctors, better follow up after diagnosis and more personalized treatment plans. [...] Thank you for considering this."
2046	"I am asking that you invest in C-PTSD (Complex PTSD) so that it can go into the DSM-5. [...] Why won't you write in a book what I have and give clear guidelines to therapists to help me? I'm right here, in front of you!"
2111	"Policymakers often face tough choices on where to allocate resources. I urge you to consider investing in research and support for individuals living with [insert illness, condition, or disability]. [...] Please, invest in us."
2941	"I have not yet found a permanent solution for my condition and I would love for there be intensive research for the condition."
4643	"As a type 1 Diabetic I firmly believe that insulin and blood sugar test strips need to be more affordable if purchasing without medical insurance."
5382	"By allocating funds to this research and public health initiative, we could potentially diagnose many more babies accurately and provide intervention"
5556	"Putting money into diabetes research and support is not only a humanitarian act, but also a wise policy one."
5922	"you could help do some good to all that has gone down the drain. [...] Please help put our poorer and suffering citizens have a beacon of hope."

6123	"To a Valued Policymaker, Policymaker, fund research for more effective treatment and accommodation of Anxiety, ADHD, and PTSD."
6874	"This is why i'masking [sic] you as policy makers, to consider investment in better treatment and accomodation [sic] options for young people my age with serious illnesses."
7050	"Dear Elected Official, I'm writing to you today not as a concerned citizen, but as someone who lives with schizophrenia. [...] Please, I urge you to prioritize mental healthcare reform."
7111	"Dealing with issues like these hardly makes life even worth living, and this is why funding should be made available to help people like myself"
7219	"There is a strong need to continue to research treatments for Rheumatoid Arthritis."
7426	"My story presents a strong case to the policymakers as to where the money should be directed in form of research and access of treatment and robust support systems, [...] When you invest in the field of innovation, and in helping chronic and life-threatening diseases, you do not merely cure the conditions. What you are investing is human potentiality."
8812	"Thank you for your convenience."
9460	"I am requesting you to no longer view people like me as a liability but as potential. We will have the opportunity to be cured, to do something important, and to live a dignified life with the proper resources at our disposal better treatment pathways, focused research, and thoughtful accommodations. Not spending a penny here is not only an act of charity; it is an investment in human capacity and social power."

Table 45: Framing statements given in Persuasive narratives.

Participant ID	Feedback
0848	"Thinking about additional funding for research made me really hopeful that there will be another treatment someday that is more effective."
1561	"I was thinking about how I should describe the illness and my points for its treatment to be funded."

2111	"I was thinking about how to balance personal experience with persuasive impact. [...] I wanted the tone to be respectful but urgent, and I aimed to emphasize both the human and economic cost of underfunding treatment and accommodations. [...] I was thinking: How do I get someone who controls resources to see that this isn't just a medical issue—it's about dignity, inclusion, and untapped potential? Ultimately, my thought process was about how to humanize the experience without making it only about struggle"
2492	"Was trying to recall all important information about the event [...] and give them in a concise and clear manner, while retaining chronological order."
5382	"I focused on how the current system failed my [RELATIVE] and I and how research into changing the system could help other patients. My goal was to write a persuasive argument for why funding should be put towards this initiative."
6213	"My thought process focused on crafting a concise, persuasive argument for policymakers, emphasizing societal benefits and economic returns for mental health research."
7426	"I organized the address so that it went through experience to policy-based plea with the necessity of treatment options and support system laid out. I aimed at demonstrating the influence of these conditions on everyday life and the need to invest in care and research."
9460	"Audience & goal: I wrote with a policymaker in mind — the aim was persuasion: show how investing in research, treatment, and accommodations yields social and economic returns. [...] Tone choices: Persuasive but measured — empathetic and personal, not overly emotional; practical and policy-minded rather than medical-technical."

Table 46: Statements given in Persuasive feedback.

Emotional Scenario:

Emotional NarraType framing statements and feedback notes are found in Tables 47 and 48.

Participant ID	Framing Statements
0534	"Hey, it's tough, I know. But you're stronger than you think. [...] remember, you're not alone in this."
1029	"I hate to see that you're going through this too but I hope my experience can help you feel as if you're informed in making your own treatment systems [...] We all deserve a life without pain, but it's okay to find joy and ways to live in the meantime while we wait for those therapies to be developed."

1616	"I know it is not easy receiving such a diagnosis. [...] You will have your why me moments but it is good to remember that you are important exactly as you are."
1816	"So, take the help that is offered to you, at all costs. [...] Do that much for yourself!"
1340	"You will need to steel yourself [...] Try to approach this with good humor, or understanding, or pity, or whatever you need"
2017	"Hi it's so nice to meet you. So we've just talked about you'r [sic] diagnosis with Major Depressive Disorder. [...] just remember, this is a journey and I'll be here to guide you and support you every step of the way."
2768	"I am here to support you even if no one else is since I know what you're going through. [...] You are your best support so take good care of yourself."
3654	"Hi there, First of all, I want to say that I am really sorry you are going through this. [...] You have got this. And I am rooting for you."
3732	"I would just like to tell you that things will be okay."
3952	"I too was diagnosed with Myasthenia Gravis."
4198	"I will be here to guide you if you ever have any questions. [...] All in all, dont lose hope."
5142	"Hey there, I know how tough it is when you first hear the news about a new diagnosis [...] You've got this, and I'm cheering you on. Take care, [NAME]"
6282	"My advice to you is simple."
7310	"We've just met, and I understand you've recently been diagnosed with depression, anxiety, and PTSD. [...] Be kind to yourself as you start to figure out your own way of dealing with these conditions and learning the best ways to take care of yourself."
7599	"Hey. That 'C' word... it's a punch to the gut, right? [...] Don't be afraid to ask for what you need. Most importantly? Hold onto hope."
8255	" You're not alone in this, and with time, you'll discover your own ways to adapt and thrive. I believe in you, and I'm rooting for you as you start this journey."
8760	"I will write about breast cancer [...] do not lose hope, and as I said if cancer is not a death sentence Thank you."
9002	"Hi. You and I are more alike than you think. [...] There is always hope, and with hope (and faith), ALL things are possible. Never give up, and especially, never give up on yourself."
9079	"I would say to the other patient that please do not give up you can still improve your condition. [...] If you want to talk about what you should do and where you start from, here is my email. We can chat and I will tell you what to do."

9101	"As someone who is suffering from Complex Post Traumatic Stress Disorder (CPTSD)as you are, I feel as though I can help. [...] Keep your head up, think positive, and I can promise you that this mental illness will be conquered!"

Table 47: Framing statements given in Emotional narratives.

Participant ID	Feedback
0534	"I wanted to offer immediate, concise encouragement. I focused on core ideas: acknowledging difficulty, highlighting strength, suggesting practical coping, and emphasizing community."
1029	"I felt like sharing a brief overview of my own migraine journey helps the person I'm talking to compare some of their overlap. But I also really wanted to focus on emphasizing the positives [...] I think new patients should know [enjoying life] is possible even for patients who don't get better"
1340	"I genuinely approached this question as if it were a speech or a letter to someone in the same situation as me."
1616	"I felt sad for the [X YEAR OLD] me who was scared and lost"
1816	"What my thought process was was to actually state it in a way in which I wish I would have been told from someone experienced in this! I actually had a friend who went through something which absolutely caused him the same diagnosis! I remember trying to help him out and explaining to him why it's so very important to seek help now rather than wait until his life is falling apart!!!"
2017	"the first thing I was thinking during the writing in question one is, I felt empathy for the patient. [...] So as I wrote my response My personal experiences were important and central. [...] I wanted to give them the support and warmth and comfort that they deserved and a genuine understanding because I remember the fear and isolation that came with my diagnosis and I wanted the patient to know that they were not alone on their journey and feelings. [...] My main goal was to give them compassion and sympathy for the challenges that they would face ahead while also giving them hope."
2768	"I was thinking about the last decade and how I felt"

3654	"I [...] tried to describe [my emotions] honestly so someone else could relate. I wanted to give encouragement without sounding fake, so I focused on what helped me most, like learning my triggers and asking for help. I also made sure to keep the tone supportive and hopeful, because I knew the person reading it might be scared or overwhelmed. My goal was to make them feel less alone and more in control of their situation."
3732	"I was trying to think empathetically and helpful during the study"
4198	"I thought of the words I wish I was told just as I found out I had the issue. I thought of how I had wanted to be encouraged."
5142	"I was really focused on how to make it feel personal and heartfelt. I wanted to offer advice and comfort in a way that would truly speak to someone who might be feeling overwhelmed [...] I tried to create a tone that felt warm and reassuring [...] honest but also hopeful [...] I wanted to make sure the person reading it felt like they weren't alone in this journey [...] I hope it comes across as real and supportive, just like I would talk to a friend who's been through the same thing."
7310	"My thought processes were focused on thinking about the most important message i wanted to convey [...] i wanted to keep it focused on advise and coping [...] I wanted to try to help in that aspect."
7599	"When I was trying to put that message together for the newbie, it was all about being real. [...] I just wanted them to know they're not the only one staring into the abyss. [...] But the main thing I wanted to send their way was a little flicker of hope."
8255	"I knew it was supposed to be for someone who just got the same diagnosis, so I wanted to make it feel real and helpful, you know? [...] I remembered how lost I felt, so I started there to set the tone. I wanted them to know it's okay to feel that way at first. [...] I was picturing this other person reading it, feeling scared, and I wanted to end on a positive note. That's why I added the bit about good days and bad days, and finding joy in little things—it's what I tell myself when it gets rough."
8585	"I was thinking of all the ways I could help highlight my journey with useful experiences that the other person could use."

Table 48: Statements given in Emotional feedback.

Clinician Scenarios:

Clinician NarraType framing statements and feedback notes are found in Tables 49 and 50.

Participant ID	Framing Statements
0243	"I'd really value your input on the optimal way to manage my diabetes in the long term, especially in how to make my treatment plan more individualized to more closely align with my life in a sustainable manner."
0448	"My previous GP helped manage my symptoms, and I'd like to continue appropriate care with you."
1498	"i made an appointment woth [sic] you [...] i need a new gp to help me with my health and medication"
2794	"Hi I'm [NAME] I'm [YEAR X] years old and have [PERSONAL DETAIL]."
4399	"Hi, and thank you for taking time to meet with me. I would like to give a general overview of my medical history as requested since this is our first appointment."
5637	"This is something very delicate about me that most of the time I choose not to share with others but since you are now my General Practitioner, there is no other means than to voice out my medical condition which has been troubling me for a very long time."
6371	"Hello, am person A and its nice to meet you. I have come for some help from you because I need help with my condition. [...] I don't have much to say but I will be happy working with you. Thank you"
8176	"My last GP diagnosed me with [...]"
8372	"I appreciate you taking the time to go over my medical history with me. Since this is my first visit with you, I'd like to give you a clear picture of how my condition has impacted my daily life."
8599	"Greetings, Doctor. Thank you for becoming my new general practitioner."
9580	"This is actually something I'm in the middle of doing, due to an insurance and job change. While I try to make sure they have some experience with my conditions, that's not always possible with GPs. I do try to ask for a longer appointment for my first visit. [...] While I can go into further detail on any of these things, I know most GPs are on a much tighter schedule than my previous private-practice GP, so I try to get the biggest things down in a condensed form."

Table 49: Framing statements given in Clinician narratives.

Participant ID	Feedback

0243	"I tried my best to imagine that I was sitting across from a new doctor and describing my own history with diabetes to him or her. I was aiming to be honest, realistic, and reflective, and not dramatic or excessively emotional. [...] I wanted it to sound personal and real, like I'm talking with someone who can help me become better in a healthier way."
0448	"I focused on keeping the response general and relevant to a first conversation with a new GP. I wanted to include key information about having a chronic disease, how it affects daily life, and the need for ongoing care."
1468	"i descrided [sic] my heavy metal [sic] dissorders [sic] like metal [sic] illness and my strong chronic disease to my gp"
1891	"I simply put myself in the situation of having to describe my history with chronic pain to a new physician who I had not seen previously."
2919	"I focused on clear memories to show how severe and disruptive the condition was."
3678	"I tried to think of the most important things I'd need to explain to a new primary care doctor, including past and current struggles and how those struggles have affected/continue to affect my life. I figured these would be important, seeing as they may handle changes in my medications."
4386	"I simply decided to share pertinent details of what currently bothers me, because why else would I be interfacing with a doctor? I wanted to include specifically what is troubling me and what I am seeking, as well as possible contributing factors to these problems."
4399	"I focused on sharing the most important parts of my health history clearly and as briefly as possible."
7454	"I was thinking [...] how frustrated I have been with the health care system during this journey."
7542	"I tried to imagine how I would speak to a new doctor in a calm, honest way, without sounding like I was complaining or overexaggerating anything."
8372	"I focused on giving a clear and honest account of how my coniditon [sic] affects my daily life [...] enough detail for a new GP to understand my challenges without overwhelming them with unnecessary information [...] make it clear that while I'm proactive, I also need support and medical advice. I tried to strike a balance between beiing [sic] descriptive and keeping my response direct and to the point. My thought process was about making my experience understandable to a medical professional in a way that would help them provide me with the best possible care."

8599	"I would need to provide a realistic and personal medical history to a new general practitioner. I began with a warm hello to establish a conversational tone, and then I focused on presenting my problems in chronological sequence to resemble a standard medical narrative [...] In order to show initiative, I brought up walking every day. I ended by asking the GP for advice in order to promote communication."
9580	"I was thinking about the way I've learned to handle doctors. Spending a lot of time on family history, or the history of my entire life doesn't actually do much good. The most important thing is to give them an idea of everything going on with me, and judge whether they're going to be helpful or not. If they're not, there's no real point [...] as I won't be continuing with them past that initial visit."

Table 50: Statements given in Clinician feedback.

Nuanced Intentions:

Table 51 shows a list of framing and feedback statements from the IlNeS corpus where the expressed intent more closely matched a distinct NarraType value. In IlNeS, these types of statements arise only in Factual and Persuasive texts–this could be due to the high rate of engagement with the Emotional NarraType prompt, and the alluded-to idea that emotional narratives are considered "default" for the illness narrative genre.

Participant ID	Prompt Given	Feedback
0135	Persuasive	"I was thinking writing about my illness gives me the opportunity to share my experience."
2276	Factual	"I also considered the emotional effect, such as being isolated or frustrated and how friends and family do not always know, even when they are supportive."
3517	Factual	"Mental illness is not taken as seriously as physical illnesses but they are equally worthy of help and care."
4677	Factual	"I was [...] thinking about how some people may not take [my condition] seriously or even believe that it is a disability."
6874	Persuasive	"i wanted to capture the heavy emotional weight that came along with the experience: the fear, the physical pain and the isolation"
8812	Persuasive	"[I] tried to organize it all into a reasonable timeline of events as I recalled them. I realized that the answers I wrote may or may not follow the exact guidelines, but I'm honestly doing my best here."

Table 51: Feedback where participants express intents outside their prompted NarraType.

10.3.3 Qualitative Tables

This section shows the full set of statements which relate to one of the following illness narrative hypotheses: 1) that talking about illness through narrative reflects a unique social experience (Table 52); 2) the impact that sharing about illness through narrative has on survey participants (Tables 53, 54, and 55); 3) that construction of illness narrative directly links to the process of recollection (Table 56); and 4) that illness narrative is often used to position the author as a moral actor in their story of illness (Table 57).

Unique Social Experience:

Participant ID	Prompt Given	Feedback
1029	Emotional	"I felt as if I was in the role of a patient advocate which is a role I have been in many times"
4643	Persuasive	"It's not everyday you're explaining your struggles to anyone because that opportunity doesn't always present itself and no one really lends an ear."
5314	Factual	"My thought process was pretty easy as i already knew what i was going to say. I've been dealing with a stutter my whole life so it's second nature"
6282	Emotional	"I've told this story several times to others, so I basically just retold it again."
9002	Emotional	"It hurts to recount them [...] But, it probably helped, too, because I literally have no one to share these things with. "
9580	Clinician	"I was thinking about the way I've learned to handle doctors."

Table 52: All participant statements which frame illness narratives as a unique experience compared to other social interactions.

Sharing Stories Leaves Impact:

Participant ID	Prompt Given	Excerpt
1616	Emotional	"Reflecting back has made me feel happy and positive"
2289	Clinician	"I felt relief as I don't have problems any more with sharing my past"
3517	Factual	"It helps me to feel seen, too."
4040	Clinician	"i was basically happy i could talk about it without tears in my eyes"
5142	Emotional	"I remember how much it meant to hear from others who understood what I was going through."
6657	Factual	"So dyslexia hasn't been a joke and I'm glad I'm able to tell another person about. it."

7599	Emotional	"Feeling all the crap ,the anger, the fear, the sadness, that's totally normal. You gotta let it out. Find someone who'll just listen without trying to fix it."
9101	Emotional	"I had a proud feeling in me. I felt proud that I overcame such a stressful time in my life, and am now discussing it with a complete stranger. I am not sure what this study is for, but I feel as though it didn't find my by chance."

Table 53: All participant statements which frame illness narratives as a positive, healing experience.

Participant ID	Prompt Given	Excerpt
0848	Persuasive	"Writing my answer to Question 1 brought up a lot of emotions and thoughts about my health journey."
2046	Persuasive	"Section 1 I was pissed, and couldn't write much. I wrote most in 2 and 3. It brought tears to my eyes, and I hope people like myself will get the recognition they deserve."
3441	Factual	"It was kind of tough reliving it, but kind of freeing at the same time."
3637	Factual	"So reminiscing of this brought back so many feelings of positive and negative emotions and made me realize how good I am that I can take something to help me"
9002	Emotional	"What I was thinking during the writing in Question 1 is how much it made me sad to write (type) about all the health issues and disabilities that have come upon me over the years. It hurts to recount them [...] But, it probably helped, too, because I literally have no one to share these things with."

Table 54: All participant statements which express mixed emotions about sharing their illness narrative.

Participant ID	Prompt Given	Excerpt
1510	Factual	"[...] this is very keen for me since it has been a pain and bad feeling for me for some years now."
2794	Clinician	"I get somewhat upset and anxious when I talk about it because people don't understand it usually so it is very frustrating"
3634	Clinician	"Recalling what really happened led to where I am right now, almost made me cry."

4643	Persuasive	"The question really made me feel vulnerable as a person and as a type 1 diabetic."
8612	Persuasive	"I cant do anymore Im sorry"
8760	Emotional	"I was feeling rather uneasy while writting [sic], I find it difficult when it comes to talking about some of the dark moments in my life."

Table 55: All participant statements which express negative feelings about sharing their illness narrative.

The impact of sharing illness narrative is, overall, more mixed than most within the field would suggest. There are multiple potential explanations for this, though the response of Participant 2046 (from Table 54) suggests one possible correlated attribute is narrative length. 2046 was assigned the Persuasive response, and answered Question 1 (the main testimonial) with a short response. In freeform feedback explaining their thought process, 2046 noted that, "I figured if someone isn't willing to help from the basic plea, it's a waste of time to explain more." But when asked to label events from the first question, 2046 instead rephrased their initial narrative, going into much more detail about the original events. In feedback, 2046 admitted that "Section 1 I was pissed, and couldn't write much. I wrote most in 2 and 3. It brought tears to my eyes." This suggests that longer narratives correlate with positive emotional catharsis–but that a participant's state of mind may limit how long a narrative they can craft.

Recollection:

Participant ID	Prompt Given	Feedback
0135	Persuasive	"As i wrote this i remembered the events of that day when i was told about my diagnosis after several tests."
1616	Emotional	"I felt taken back to the position of the patient who just got diagnosed."
2451	Factual	"I struggle with remembering what happened during that time."
2492	Persuasive	"Was trying to recall all important information about the event."
2919	Clinician	"I tried to reflect on the specific moments when the allergy first started and how it changed my day-to-day life."
3441	Factual	"I tried reliving from the very start."
3634	Clinician	"As I was writing this piece, it enabled me to think through the painful memories that I had in my life. Recalling what really happened led to where I am right now"
3654	Emotional	"I tried to put myself back in the moment when I was first dealing with my condition."
3837	Factual	"As I wrote the story, I reminisced all of the times that I felt so terrible and left behind and I had no confidence in my abilities in myself."
4073	Factual	"I was thinking about how this experience has spanned over a decade of my life."

5056	Factual	"I just was focused on trying to recall the experience I had. How it felt and how it happened."
5345	Factual	"I was mostly trying to remember what happened back then. I still have a little brain fog, not only from COVID and Long COVID, but also since I'm entering menopause."
6536	Persuasive	"I just really wanted to try to give a cohesive picture and timeline of how the whole process went, because sometimes it was a real blur [...] I really have to kind of view it as a series of events or a story just to keep myself from getting anxious and disgusted with myself."
7070	Persuasive	"I thought about how my life was before having schizophrenia."
7135	Clinician	"It brought back a lot of memories thinking about when my health issues started."
8196	Clinician	"I was thinking about the look on my dad's face as I turned blue and that he had a cigarette in his hand. [...] I can almost smell his cigarettes. Memories are freaking weird."
8660	Clinician	"I was replaying the past year in regard to my heart disease diagnosis."

Table 56: Collection of excerpts which describe the process of recollection.

Moral Element:

ID	Prompt	Condition	Narrative
Positioning Self as Not At Fault			
0848	Persuasive	Celiac disease	"[...] even trying to eat perfectly, I still feel sick more often that I would like."
1334	Persuasive	Not named	"I had dome [sic] everything that I was supposed to do, in the order prescribed, but one day, none of that mattered."
4578	Factual	Not named	"I'd sometimes get the feeling people thought I was just being lazy or making excuses. [...] I wasn't trying to be difficult or avoid doing things"
4643	Persuasive	Diabetes	"It's not fair and no one asked to be born sick or develop deadly diseases and to just throw their lives away."
8196	Factual	Dyslexia	"If I could tell people one thing about dyslexia, it would be this: It's not about laziness or a lack of intelligence."
9002	Emotional	Lung issues	"I developed lung issues, God only knows how. I've never smoked a day in my life"

Positioning Self as Responding Reasonably			
0243	Clinician	Diabetes	"I was aiming to be honest, realistic, and reflective, and not dramatic or excessively emotional."
2111	Persuasive	Not named	"We're not asking for pity", "I was careful to avoid sounding like I was asking for sympathy."
4578	Factual	[Fatigue]	"I didn't want to sound like I was making excuses."
7542	Clinician	Arthritis	"I tried to imagine how I would speak to a new doctor in a calm, honest way, without sounding like I was complaining or overexaggerating anything."
Positioning Self as Taking Responsibility			
4040	Clinician	Obesity, depression	"I had bad eating habits and I barely took care of myself, in other not to feel bad about how I treated my body, I would normally encourage people around me not to take care of their health and physical look, I used to tell them that people need to accept them and take them as they were."
5637	Clinician	Depression, anxiety	"I feel so empty and ashamed of myself a lot of most of the time [...] I really wished I can get all the needed help I could to help me with this my condition because to be very honest, I feel so ashamed of myself a lot and wished my life is on a different level."
6536	Persuasive	Idiopathic intracranial hypertension	"I just feel like I brought it on myself."
8812	Persuasive	Pulmonary artery disease	"The quality of life is very poor and I realize I have no one to blame other than myself. [...] If there were a cheaper way to get these treatments, I would gladly reconcile with my past mistakes and hope to lead a better life in the long run."
9079	Emotional	Not named	"This condition happens to a lot of people and it is due to our decisions and a sedentary lifestyle."

Table 57: Collection of excerpts dealing with moral elements of illness narrative.

10.4 Appendix IV: TempR-MInt Experiments

This section of the appendix covers additional information from the TempR-MInt experiments.

10.4.1 Additional IlNeS data

Final IlNeS Dataset:

Table 58 shows the distribution of labels in the final IlNeS, randomizes the ordering of events within an event-pair compared to the raw IlNeS* (i.e. for a pair $Pair(E_1, E_2)$ from IlNeS* such that E_1 appears before E_2 in the text, the event-pair may be inserted into IlNeS either as $Rel(E_1, E_2) = label$ or as $Rel(E_2, E_1) = reverselabel$, where $reverselabel = Reverse(label)$). It also counts the specific grouping of the following label types: no-overlap ($before$ and $after$), full-overlap ($simultaneous$), and partial-overlap ($includes$ and $isincluded$). Due to the properties of TEO/ETRE labels, these distributions remain the same before and after the randomization process.

Label	Factual	Persuasive	Emotional	Clinician	Total
Before	.425	.329	.458	.383	.410
After	.461	.309	.446	.377	.417
Simultaneous	.035	.059	.075	.033	.049
Includes	.034	.143	.013	.106	.059
Is Included	.045	.159	.008	.100	.064
No-overlap	.887	.638	.904	.760	.828
Full-overlap	.035	.059	.075	.033	.049
Partial-overlap	.079	.303	.021	.206	.123

Table 58: The distribution of labels in the final IlNeS dataset, which randomizes input pair order to ensure TempR-MInt models learn to differentiate between label types.

For this distribution, a majority classifier is ill-suited to the task. A simple majority given this data will predict $Rel(E_X, E_Y) = after$ for all $Pair(E_X, E_Y)$, and output F1 of **24.672**. TempR-MInt began experiments with a simple implementation of MulCo using basic parameter tuning and no label smoothing. This version of the model made predictions for both the $before$ and $after$ label and achieved F1 of **45.214**. It therefore greatly outperformed simple majority classification; however, it was important to determine if this version of the model successfully learned features which could differentiate between the two most common $before$ and $after$ labels.

To test this, TempR-MInt compared the performance of MulCo against a model whose predictive formula could be conceptualized as the following:

$$Rel(E_X, E_Y) = Rand([before, after]) \tag{21}$$

A model using this predictive formula achieves an a F1 of **41.393**. MulCo improves on this score by nearly 4 points, which demonstrates that even the initial MulCo implementation extracts some meaningful predictive data compared to random selection.

MulCoS with Label Smoothing

Table 59 shows the confusion matrix of the MulCo-sensitive model (or MulCoS) with label smoothing on the final IlNeS dataset.

Predicted Label → Gold Label	Before	After	Simultaneous	Includes	Is Included	Total
Before	**2007**	214	6	24	6	2257
After	207	**2062**	6	0	22	2297
Simultaneous	125	118	**14**	0	7	264
Includes	246	69	3	**6**	2	326
Is Included	66	270	5	0	**11**	352
Total	2651	2733	34	30	48	5496

Table 59: Confusion matrix for MulCo-sensitive on the IlNeS test dataset.

Performance of TEO/ETRE models has historically benefited from a slight bias towards a majority classifier, and therefore models often perform best on datasets with significant prior class bias. Models trained on IlNeS* (which has a strong bias towards the $before$ label) benefit from this much more than models trained on IlNeS (with two roughly-equivalent majority labels). There is a sharp drop in performance on the IlNeS dataset compared to IlNeS*, regardless of other factors. However, through use of order-sensitive MulCoS, the version of the model trained on IlNeS* is fully competitive with the simpler majority classification on IlNeS* (with a score of **66.110** compared to the IlNeS* majority's **64.341**). The implementation of MulCoS allows a more complicated, meaningful dataset to be used as a comparative baseline for the TempR-MInt experiments.

Other benefits of the MulCoS/IlNeS approach can be extracted from the confusion matrix for predictions (see Table 59). Qualitative improvements achieved by the MulCoS/IlNeS model do not necessarily lead to improvements in F1 score, but represent real progress in the task. For one, this model pairs well with label smoothing, and makes at least some correct predictions for all label types. Further, the most common errors this model made was mistaking *includes* pairs for $before$, and *isincluded* for $after$. These pairs of labels share the property *chronological directionality* (i.e. in both $Rel(E_A, E_B) = before$ and $Rel(E_A, E_B) = includes$, E_A begins before the start of E_B). That the model confuses semantically-similar temporal labels demonstrates an understanding of time that still leads to "incorrect" predictions, but is a clear stepping stone to full understanding of the TEO/ETRE task.

10.4.2 LIWC Tables

The following section shows the mean values for frequency of LIWC dictionaries in each testimonial per NarraType, and marks where statistically-significant distinctions exist. Each table (Tables 60, 61, 62) lists LIWC category names along with the category's parent dictionary. Note that the hierarchical nesting of LIWC dictionaries has a total depth of 3.

The NarraType partitions display statistically-significant differences in LIWC frequency values for numerous dictionaries. These results inform the design of TempR-MInt experiments which use LIWC values as proxies for NarraType, as detailed in Section 8.3.3.

LIWC category	Factual	Persuasive	Emotional	Clinician	p-value
Health (Physical)	3.914^{ab}	5.626^{a}	2.871^{b}	5.351^{a}	.0001
Illness (Health)	1.603^{bc}	2.07^{ab}	$.98^{c}$	2.403^{a}	.008
Lifestyle (N/A)	1.711^{b}	3.609^{a}	1.088^{b}	1.503^{b}	7.512e-7
Money (Lifestyle)	$.124^{b}$	1.4^{a}	$.148^{b}$	$.131^{b}$	2.138e-5
Physical (N/A)	5.432^{ab}	7.385^{a}	4.208^{b}	7.507^{a}	.0001
SocBehav (Social)	2.178^{b}	2.565^{ab}	3.371^{a}	2.034^{b}	.001
SocRefs (Social)	2.709^{b}	4.261^{b}	6.457^{a}	2.07^{b}	1.53e-6
Want (States)	$.132^{b}$	$.112^{b}$	$.331^{a}$	$.268^{ab}$	.007
Work (Lifestyle)	1.028^{b}	2.281^{a}	$.717^{b}$	1.174^{ab}	.0002

Table 60: LIWC categories related to topics with significant distinctions between NarraType. We use letters per row to mark which datasets differ significantly. Entries in a row which do not share a letter have statistically-significant differences in frequency.

LIWC category	Factual	Persuasive	Emotional	Clinician	p-value
Allure (Motives)	7.802^{b}	6.26^{c}	8.854^{a}	7.6^{b}	3.51e-5
Analytic (N/A)	29.748^{b}	52.042^{a}	27.044^{b}	41.44^{ab}	.0003
Assent (Conversation)	$.071^{ab}$	$.017^{b}$	$.248^{a}$	$.049^{b}$	.003
Authentic (N/A)	83.174^{ab}	57.976^{c}	70.206^{bc}	92.226^{a}	2.46e-5
BigWords (N/A)	16.052^{b}	23.227^{a}	14.539^{b}	17.12^{b}	1.33e-7
Clout (N/A)	8.707^{b}	23.49^{a}	40.314^{a}	4.531^{b}	4.188e-7
Conversation (N/A)	$.148^{ab}$	$.046^{b}$	$.429^{a}$	$.056^{b}$	.002
Insight (CogProc)	3.261^{ab}	2.711^{ab}	4.356^{a}	1.946^{b}	.0001
Linguistic (N/A)	$70.811^{a}b$	65.396^{c}	72.014^{a}	$67.600^{b}c$	.0002
Negate (Function)	1.685^{a}	1.416^{ab}	1.92^{ab}	$.873^{b}$	.003
Prosocial (SocBehav)	$.419^{b}$	1.425^{a}	1.314^{a}	$.700^{ab}$	.0003
Social (N/A)	$5.404^{b}c$	$7.49^{a}b$	10.604^{a}	4.753^{c}	4.188e-7
Tone (N/A)	17.747^{b}	29.593^{ab}	40.918^{a}	25.555^{ab}	.007
TonePos (Affect)	2.088^{b}	3.048^{ab}	3.59^{a}	2.345^{b}	.0009

Table 61: LIWC categories related to rhetoric with significant distinctions between NarraType.

LIWC category	Factual	Persuasive	Emotional	Clinician	p-value
FocusFuture (Perception)	1.576^{a}	$.948^{ab}$	1.924^{a}	$.691^{b}$	.003
FocusPresent (Perception)	4.565^{ab}	4.437^{ab}	5.433^{a}	3.340^{b}	.0007
Time (Perception)	5.82^{ab}	4.961^{ab}	4.673^{b}	7.579^{a}	.0002

Table 62: LIWC categories related to time with significant distinctions between NarraType.

References

Abu-Jbara, Amjad, Jefferson Ezra, and Dragomir Radev (2013). "Purpose and Polarity of Citation: Towards NLP-Based Bibliometrics." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 596–606.

Allen, James F. (1983). "Maintaining Knowledge about Temporal Intervals". In: *Communications of the ACM* 26.11, pp. 832–843.

— (1984). "Towards a General Theory of Action and Time". In: *Artificial Intelligence* 23.2, pp. 123–154.

— (1991). "Time and Time Again: The Many Ways to Represent Time". In: *International Journal of Intelligent Systems* 6.4, pp. 341–355.

Allen, James F. and Patrick J. Hayes (1985). "A Common-Sense Theory of Time". In: *International Joint Conference on Artificial Intelligence* 85, pp. 528–531.

Amodei, Dario (2025). *The Urgency of Interpretability*. URL: `https://www.darioamodei.com/post/the-urgency-of-interpretability` (visited on 04/2025).

APA, American Psychological Association (2017a). *PTSD Guidelines - Treatments - Cognitive Behavioral Therapy*. URL: `https://www.apa.org/ptsd-guideline/treatments/cognitive-behavioral-therapy` (visited on 07/31/2017).

— (2017b). *PTSD Guidelines - Treatments - Narrative Exposure Therapy*. URL: `https://www.apa.org/ptsd-guideline/treatments/narrative-exposure-therapy` (visited on 07/31/2017).

Atkinson, Paul (2009). "Illness Narratives Revisited: The Failure of Narrative Reductionism". In: *Sociological Research Online* 14.5, pp. 196–205.

atundratoadstool (2022). URL: `https://atundratoadstool.tumblr.com/post/683528925666770944/i-just-want-everyone-new-to-dracula-and-reading` (visited on 05/06/2022).

Ballesteros, Miguel, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen Mckeown, and Yaser Al-Onaizan (2020). "Severing the Edge Between Before and After: Neural Architectures for Temporal Ordering of Events". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5412–5417.

Bamman, David, Sabrina Baur, Mackenzie Hanh Cramer, Anna Ho, and Tom McEnaney (2025). "Measuring the Stories in Contemporary Songs". In: *Poetics* 13, p. 826.

Bell, Allan (1984). "Language Style as Audience Design". In: *Language in Society* 13, pp. 145–204. DOI: `10.1017/S004740450001037X`.

Beltagy, Iz, Matthew E Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: *arXiv preprint arXiv:2004.05150*.

Beniwal, Himanshu, Dishant Patel, Hritik Ladia, Ankit Yadav, and Mayank Singh (2024). "Remember This Event That Year? Assessing Temporal Information and Reasoning in Large Language Models." In: *arXiv preprint arXiv:2402.11997*.

Bhatia, Parminder, Yangfeng Ji, and Jacob Eisenstein (Sept. 2015). "Better Document-level Sentiment Analysis from RST Discourse Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, p. 2212. DOI: `10.18653/v1/D15-1263`. URL: `https://aclanthology.org/D15-1263/`.

Boguraev, Branimir, James Pustejovsky, Rie Ando, and Marc Verhagen (2007). "TimeBank Evolution as a Community Resource for TimeML Parsing." In: *Language Resources and Evaluation* 41.1, pp. 91–115.

Boyd, Ryan L., Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker (2022). "The Development and Psychometric Properties of LIWC-22". In: *Austin, TX: University of Texas at Austin* 10, pp. 1–47.

Boyd, Ryan L. and Andrew H. Schwartz (2021). "Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field". In: *Journal of Language and Social Psychology* 40, pp. 21–41.

Bramsen, Philip, Pawan Deshpande, Yoon Keok Lee, and Regina Barzilay (2006). "Inducing Temporal Graphs". In: *In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 189–198.

Breitfeller, Luke, Aakanksha Naik, and Carolyn Rosé (2021). "STAGE: Tool for Automated Extraction of Semantic Time Cues to Enrich Neural Temporal Ordering Models". In: *arXiv preprint arXiv:2105.07314*.

Buckner, Randy L. and Mark E Wheeler (2001). "The Cognitive Neuroscience of Remembering". In: *Nature Reviews Neuroscience* 2, pp. 624–634.

Bury, Mike (2001). "Illness Narratives: Fact or Fiction?" In: *Sociology of Health & Illness* 23, pp. 263–285.

Byrne, Alex (2010). "Recollection, Perception, Imagination". In: *Philosophical Studies* 148, pp. 15–26.

Camara, Bienvenu Salim, Loubna Belaid, Hawa Manet, Delphin Kolie, Etienne Guillard, Théophile Bigirimana, and Alexandre Delamou (2020). "What Do We Know abbout Patient-Provider Interactions in Sub-Saharan Africa? A Scoping Review". In: *Pan African Medical Journal* 37.

Cassidy, Taylor, Nathanael Chambers, Bill McDowell, and Steven Bethard (2014). "An Annotation Framework for Dense Event Ordering". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 2: Short Papers, pp. 501–506.

Chambers, Nathanael, Taylor Cassidy, Bill McDowell, and Steven Bethard (2014). "Dense Event Ordering with a Multi-Pass Architecture". In: *Transactions of the Association for Computational Linguistics*. Vol. 2, pp. 273–284.

Chambers, Nathanael and Dan Jurafsky (2009). "Unsupervised Learning of Narrative Schemas and Their Participants". In: *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 602–610.

Chang, Angel X. and Christopher D. Manning (2012). "SUTime: A Library for Recognizing and Normalizing Time Expressions". In: *Lrec* 12, pp. 3735–3740.

Charon, Rita (2008). *Narrative Medicine: Honoring the Stories of Illness*. Oxford University Press.

Chen, Shuang, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu (2025). "Perceive the Passage of Time: A Systematic Evaluation of Large Language Model in Temporal Relativity". In: *In Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8304–8313.

Chen, Xinlei and Kaiming He (2021). "Exploring Simple Siamese Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758.

Cheng, Fei and Yusuke Miyao (2017). "Classifying Temporal Relations by Bidirectional LSTM over Dependency Paths". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–6.

Chu, Zheng, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin (2023). "TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models". In: *arXiv*. DOI: `arXiv:2311.17667`.

Cicchetti, Domenic V. and Sara A. Sparrow (1981). "Developing Criteria for Establishing Interrater Reliability of Specific Items: Applications to Assessment of Adaptive Behavior". In: *American Journal of Mental Deficiency* 86, pp. 127–137.

Daumé III, Hal (2007). "Frustratingly Easy Domain Adaptation". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263.

Davidson, Donald (1969). "The Individuation of Events". In: *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday*, pp. 216–234.

Delbanco, Tom, Jan Walker, Sigall K. Bell, Jonathan D. Darer, Joann G. Elmore, Nadine Faran, Henry J. Feldman, James D. Ralston, Stephen E. Ross, R. Mejilla, and L. Ngo (2012). "Inviting Patients to Read Their Doctors' Notes: A Quasi-Experimental Study and a Look Ahead". In: *Annals of Internal Medicine* 157.7, pp. 461–470.

Delbanco, Tom, Jan Walker, Jonathan D. Darer, Joann G. Elmore, Henry J. Feldman, Suzanne G. Leveille, James D. Ralston, Stephen E. Ross, Elisabeth Vodicka, and Valerie D. Weber (2010). "Open Notes: Doctors and Patients Signing On". In: *Annals of Internal Medicine* 153.2, pp. 121–125.

Dinika, Adio (Sept. 25, 2024). "The Human Cost Of Our AI-Driven Future". In: *Noema*. URL: `https://www.noemamag.com/the-human-cost-of-our-ai-driven-future/` (visited on 09/25/2024).

Donovan, Jenny L. and David R. Blake (1992). "Patient Non-Compliance: Deviance or Reasoned Decision-Making?" In: *Social Science & Medicine* 34.5, pp. 507–513.

Dutt, Ritam, Zhen Wu, Jiaxin Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Rosé (2024). "Leveraging Machine-Generated Rationales to Facilitate Social Meaning Detection in Conversations". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1: Long Papers, pp. 6901–6929.

Dwork, Cynthia, Ravi Kumar, Moni Naor, and Dandapani Sivakumar (2001). "Rank Aggregation Methods for the Web". In: *In Proceedings of the 10th international conference on World Wide Web*, pp. 613–622.

EpiBio, NIH CC RMD Epidemiology & Biostatistics Section (2025a). "Annotation Guideline for Free-Text Activity Functioning Information: Communication & Cognition (1.1)". In: *Zenodo*. DOI: `10.5281/zenodo.15115291`.

— (2025b). "Annotation Guideline for Free-Text Activity Functioning Information: Mobility (1.1)". In: *Zenodo*. DOI: `10.5281/zenodo.15101601`.

Fligner, Michal A. and Joseph S. Verducci (1986). "Distance Based Ranking Models." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 48, pp. 359–369.

Fontaine, André and William A. Glavin (1987). *The Art of Writing Nonfiction*. Syracuse University Press.

Frank, Arthur W. (2013). *The Wounded Storyteller: Body, Illness & Ethics, Second Edition*. Chicago: University of Chicago Press.

Fried, Linda P., Luigi Ferrucci, Jonathan Darer, Jeff D. Williamson, and Gerard Anderson (2004). "Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications for Improved Targeting and Care". In: *The Journals of Gerontology: Series A* 59, pp. M255–M263.

Friedman, William J. (1993). "Memory for the time of past events". In: *Psychological bulletin* 113.1, p. 44.

Funnell, Martha Mitchell and Robert M. Anderson (2000). "The Problem with Compliance in Diabetes". In: *JaMa* 284.13, p. 1709.

Genette, Gérard (1980). *Narrative Discourse: An Essay in Method*. Trans. by Cornell University Press. Cornell University Press.

Grünloh, Christiane, Hanife Rexhepi, Åsa Cajander, Rose-Mharie Åhlfeldt, Gunilla Myreteg, and Isto Huvila (2016). "Patient Empowerment Meets Concerns for Patients: A Study of Patient Accessible Electronic Health Records in Sweden". In: *Joint conference that united the Medical Informatics Europe (MIE) conference, the conferences of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS), the German Society for Epidemiology (DGEpi), the International Epidemiological Association–European Region and the European Federation for Medical Informatics (EFMI)*.

Hall, Judith A., Debra L. Roter, and Cynthia S. Rand (1981). "Communication of Affect between Patient and Physician." In: *Journal of Health and Social Behavior* 22, pp. 18–30. DOI: 10.2307/2136365.

Hall-Law, Lauren, Claire Cowie, Catherine Lai, Nina Markl, Stephen Joseph McNulty, Shan-Jan Sarah Liu, Clare Llewellyn, Beatrice Alex, Elliott Zuzana, and Anita Klingler (2022). "The Lothian Diary Project: Sociolinguistic Methods During the COVID-19 Lockdown". In: *Linguistics Vanguard* 8, pp. 321–330. DOI: 10.1515/lingvan-2021-0053.

Han, Rujun, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng (2019a). "Deep Structured Neural Network for Event Temporal Relation Extraction". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 666–106.

Han, Rujun, Qiang Ning, and Nanyun Peng (2019b). "Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 434–444.

Hassol, Andrea, James M. Walker, David Kidder, Kim Rokita, David Young, Steven Pierdon, Deborah Deitz, Sarah Kuck, and Eduardo Ortiz (2004). "Patient Experiences and Attitudes about Access to a Patient Electronic Health Care Record and Linked Web Messaging". In: *Journal of the American Medical Informatics Association* 11.6, pp. 505–513.

Heikkilä, Melissa (Feb. 21, 2023). "How OpenAI is Trying to Make ChatGPT Safer and Less Biased". In: *MIT Technology Review*. URL: https://www.technologyreview.com/2023/02/21/1068893/how-openai-is-trying-to-make-chatgpt-safer-and-less-biased/ (visited on 02/21/2023).

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015). "Distilling the Knowledge in a Neural Network". In: *arXiv preprint arXiv:1503.02531* 2.7.

Huang, Quzhe, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao (2023). "More than Classification: A Unified Framework for Event Temporal Relation Extraction". In: *arXiv preprint arXiv:2305.17607*.

Hydén, Lars-Christer (1997). "Illness and Narrative". In: *Sociology of Health & Illness* 19.1, pp. 48–69.

Irurozki, Ekhine, Borja Calvo, and Josa A. Lozano (2016). "PerMallows: An R Package for Mallows and Generalized Mallows Models." In: *Journal of Statistical Software* 71.6, pp. 1–30.

Ji, Yangfeng and Jacob Eisenstein (2014). "Representation Learning for Text-Level Discourse Parsing". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* 1, pp. 13–24.

Jia, Zhen, Abdalghani Abujabal, Risharaj Saha Roy, Jannik Strötgen, and Gerhard Weikum (2018). "TempQuestions: A Benchmark for Temporal Question Answering". In: *Companion Proceedings of the The Web Conference*, pp. 1057–1062.

Jilka, Sagar Ramesh, Ryan Callahan, Nick Sevdalis, Erik K. Mayer, and Ara Darzi (2015). ""Nothing About Me Without Me": An Interpretative Review of Patient Accessible Electronic Health Records". In: *Journal of Medical Internet Research* 17.6, e161.

Kelter, Stephanie, Barbara Kaup, and Berry Claus (2004). "Representing a Described Sequence of Events: A Dynamic View of Narrative Comprehension". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.2, p. 451.

Kirkland, Matt (2021). *Dracula Daily*. URL: https://draculadaily.substack.com/.

Kitaev, Nikita, Lukasz Kaiser, and Anselm Levskaya (2020). "Reformer: The Efficient Transformer". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rkgNKkHtvB.

Korsch, Barbara M. and Vida Francis Negrete (1972). "Doctor-Patient Communication". In: *Scientific American* 227.2, pp. 66–75.

Kruk, Julia, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran (2019). "Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts". In: *arXiv preprint* arXiv:1904.09073.

Kwame, Abukari and Pammla M. Petrucka (2021). "A Literature-Based Study of Patient-Centered Care and Communication in Nurse-Patient Interactions: Barriers, Facilitators, and the Way Forward". In: *BMC Nursing* 20, p. 158. DOI: 10.1186/s12912-021-00684-2.

Labov, William (2013). *The Language of Life and Death: The Transformation of Experience in Oral Narrative*. Cambridge: Cambridge University Press.

Labov, William and Joshua Waletzky (1997). "Narrative Analysis, Oral Versions of Personal Experience". In: *Journal of Narrative and Life History* 7, pp. 3–38. DOI: 10.1075/jnlh.7.02nar.

Lapata, Mirella and Alex Lascarides (2006). "Learning Sentence-Internal Temporal Relations". In: *Journal of Artificial Intelligence Research* 27, pp. 85–117.

Leonardi, Matilde, Jerome Bickenbach, Tevfik B. Ustun, Nenad Kostanjsek, and Somnath Chatterji (2006). "The Definition of Disability: What is in a Name?" In: *The Lancet* 368, pp. 1219–1221.

Levelt, Willem J.M. (1983). "Monitoring and Self-Repair in Speech". In: *Cognition* 14.1, pp. 41–104. DOI: https://doi.org/10.1016/0010-0277(83)90026-4.

Li, Kang, Lequan Yu, Shujun Wang, and Pheng-Ann Heng (2020). "Towards Cross-Modality Medical Image Segmentation with Online Mutual Knowledge Distillation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01, pp. 775–783.

Liu, Jian, Jinan Xu, Yufeng Chen, and Yujie Zhang (2021). "Discourse-Level Event Temporal Ordering with Uncertainty-Guided Graph Completion". In: *IJCAI*, pp. 3871–3877.

Llorens, Hector, Estela Saquete, and Borga Navarro (2010). "TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2". In: *Proceedings of the 5th International Workshop on Semantic Evaluation* 284–291.

Lu, Jing and Vincent Ng (2017). "Joint Learning for Event Coreference Resolution". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vol. 1: Long Papers, pp. 90–101.

Man, Hieu, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen (2022). "Selecting Optimal Context Sentences for Event-Event Relation Extraction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10, pp. 11058–11066.

Mani, Inderjeet, Marc Verhagen, Ben Wellner, Lee Chugmin, and James Pustejovsky (2006). "Machine Learning of Temporal Relations". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 753–760.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Stever J. Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 55–60.

Mathur, Puneet, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha (2021). "TIMERS: Document-Level Temporal Relation Extraction". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 524–533.

Metz, Christian (1991). *Film Language: A Semiotics of the Cinema*. Trans. by University of Chicago Press. University of Chicago Press.

Mitamura, Teruko, Zhengzhong Liu, and Eduard H. Hovy (2017). "Events Detection, Coreference and Sequencing: What's Next? Overview of the TAC KBP 2017 Event Track." In: *TAC*.

Naik, Aakanksha, Luke Breitfeller, and Carolyn Rosé (2019). "TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events"". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 239–249.

Neumann, Arne (2015). "discoursegraphs: A Graph-Based Merging Tool and Converter for Multilayer Annotated Corpora". In: *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pp. 309–312.

Nietfeld, Emi (Dec. 2024). "What the Most Famous Book About Trauma Gets Wrong". In: *Mother Jones*. URL: `https://www.motherjones.com/media/2024/12/trauma-body-keeps-the-score-van-der-kolk-psychology-therapy-ptsd/` (visited on 12/2024).

Ning, Qiang, Zhili Feng, and Dan Roth (2017). "A Structured Learning Approach to Temporal Relation Extraction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1027–1037.

Ning, Qiang, Hao Wu, and Dan Roth (2018). "A Multi-Axis Annotation Scheme for Event Temporal Relations". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1318–1328.

Ong Lucille M.L., De Haes, Johanna C.J.M., Alaysia M. Hoos, and Frits B. Lammes (1995). "Doctor-Patient Communication: A Review of the Literature." In: *Social science & medicine* 40.7, pp. 903–918.

OSHA, Occupational Health & Safety Administration (2021). *Critical Incident Stress Guide*. URL: `https://www.osha.gov/emergency-preparedness/guides/critical-incident-stress` (visited on 09/21/2021).

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Peng, Haoruo, Yangqiu Song, and Dan Roth (2016). "Event Detection and Co-Reference with Minimal Supervision". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 392–402.

Pfeiffer, David (1999). "The Problem of Disability Definition: Again". In: *Disability and Rehabilitation* 21.8, pp. 392–395. DOI: `10.1080/096382899297530`.

Piper, Andrew and Sunyam Bagga (2022). "Toward a Data-Driven Theory of Narrativity". In: *Literary History* 54.1, pp. 879–901.

Piper, Andrew, Richard Jean So, and David Bamman (2021). "Narrative Theory for Computational Narrative Understanding." In: *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 298–311.

Pustejovsky, James, José M. Castano, Rober Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz Setzer, and Dragomir R. Radev (2003a). "TimeML: Robust Specification of Event and Temporal Expressions in Text". In: *New Directions in Question Answering* 3, pp. 28–34.

Pustejovsky, James, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, and Dragomir Radev (2003b). "The TimeBank Corpus". In: *Corpus linguistics*, p. 40.

Pyper, Cecilia, Justin Amery, Marion Watson, and Claire Crook (2004). "Access to Electronic Health Records in Primary Care-A Survey of Patients' Views". In: *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 10.11, SR17–22.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *In Association for Computational Linguistics (ACL) System Demonstrations*.

Quine, Willard Van Orman (1985). "Events and Reification." In: *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, pp. 162–171.

Reimers, Nils, Nazanin Dehghani, and Iryna Gurevych (2016). "Temporal Anchoring of Events for the TimeBank Corpus". In: *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 1, pp. 2195–2204.

— (2018). "Event Time Extraction with a Decision Tree of Neural Classifiers". In: *Transactions of the Association for Computational Linguistics* 6, pp. 77–89.

Rini, Christine, Jane Austin, Lisa M. Wu, Gary Winkel, Heiddis Valdimarsdottir, Annette L. Stanton, Luis Isola, Scott Rowley, and William H. Redd (2014). "Harnessing Benefits of Helping Others: A Randomized Controlled Trial Testing Expressive Helping to Address Survivorship Problems After Hematopoietic Stem Cell Transplant". In: *Health Psychology* 33, p. 1541.

Rogers, Anna, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky (2024). "NarrativeTime: Dense Temporal Annotation on a Timeline". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12053–12073. URL: https://aclanthology.org/2024.lrec-main.1054/.

Rose, Suzanna and Jonathan Bisson (1998). "Brief Early Psychological Interventions Following Trauma: A Systematic Review of the Literature". In: *Journal of Traumatic Stress* 11, pp. 697–710. DOI: 10.1023/A:1024441315913.

Ross, Stephen E., Jamie Todd, Laurie A. Moore, Brenda L. Beaty, Loretta Wittevrongel, and Chen-Tan Lin (2005). "Expectations of Patients and Physicians Regarding Patient-Accessible Medical Records". In: *Journal of Medical Internet Research* 7.2, e399.

Sap, Maarten, Ronan LeBras, Daniel Fried, and Yejin Choi (2022). "Neural Theory-Of-Mind? On the Limits of Social Intelligence in Large LMs." In: *arXiv*. DOI: arXiv:2210.13312.

Schank, Roger C. and Robert P. Abelson (1975). "Scripts, Plans, Goals, and Understanding". In: *In Proceedings of the 4th International Joint Conference on Artificial Intelligence* 1.

Seed, David (1985). "The Narrative Method of Dracula". In: *Nineteenth-Century Fiction* 40.1, pp. 61–75. DOI: 10.2307/3044836..

Showalter, Elaine (2002). *Teaching Literature*. Wiley-Blackwell. ISBN: 0631226249.

Sneller, Betsy and Adam Barnhardt (2023). "Sociolinguistic Prompts in the 21st Century: Uniting Past Approaches and Current Directions". In: *Language and Linguistics Compass* 17.3.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Joe Ellis, Teruko Mitamura, Hoa Trang Dang, Yukari Yamakawa, and Sue Holm (2016). "Event Nugget and Event Coreference Annotation." In: *In Proceedings of the Fourth Workshop on Events*, pp. 37–45.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma (2015). "From Light to Rich ERE: Annotation of Entities, Relations, and Events". In: *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 89–98.

Strawson, Galen (2004). "Against Narrativity". In: *Ratio* 17.4, pp. 428–452.

Strotgen, Jannik and Michael Gertz (2010). "HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions". In: *Proceedings of the 5th International workshop on semantic evaluation*, pp. 321–324.

Sweeney, Latanya (2000). "Simple Demographics Often Identify People Uniquely". In: *Health (San Francisco)* 671, pp. 1–34.

Teutsch, Carol (2003). "Patient–Doctor Communication". In: *Medical Clinics* 87.5, pp. 1115–1145.

Tian, Yonglong, Dilip Krishnan, and Phillip Isola (2020). "Contrastive Representation Distillation". In: *International Conference on Learning Representations*.

Tieu, Lina, Urmimala Sarkar, Dean Schillinger, James D. Ralston, Neda Ratanawongsa, Rena Pasick, and Courtney R. Lyles (2015). "Barriers and Facilitators to Online Portal Use Among Patients and Caregivers in a Safety Net Health Care System: A Qualitative Study". In: *Journal of Medical Internet Research* 17.12, e275.

Tulving, Endel (2002). "Episodic Memory: From Mind to Brain". In: *Annual Review of Psychology* 53.1, pp. 1–25.

Usita, Paula M., Ira E. Hyman, and Keith C. Herman (1998). "Narrative Intentions: Listening to Life Stories in Alzheimer's Disease". In: *Journal of Aging Studies* 12.2, pp. 185–197.

UzZaman, Naushad and James Allen (2010). "TRIPS and TRIOS system for TempEval-2: Extracting Temporal Information from Text." In: *Proceedings of the 5th International Workshop on Semantic Evaluation* 2: Proceedings of the Sev- enth International Workshop on Semantic Evaluation, pp. 276–283.

UzZaman, Naushad, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pusejovsky (2013). "SemEval-2013 Task 1: Tempeval-3: Evaluating Time Expressions, Events, and Temporal Relations." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM)* 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, pp. 1–9.

Van der Kolk, Bessel A. (1994). "The Body Keeps The Score: Memory and the Evolving Psychobiology of Posttraumatic Stress". In: *Harvard Review of Psychiatry* 1, pp. 253–265.

— (1998). "Trauma and Memory". In: *Psychiatry and Clinical Neurosciences* 52, S52–S64.

Vygotsky, Lev (1986). *Thought and Language*. Ed. and trans. by Alex Kozulin. Cambridge: The Massachusetts Institute of Technology.

Wang, Haoyu, Muhao Chen, Hongming Zhang, and Dan Roth (2020). "Joint Constrained Learning for Event-Event Relation Extraction". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 696–706.

Wang, Yuqing and Yun Zhao (2023). "TRAM: Benchmarking Temporal Reasoning for Large Language Models". In: *arXiv preprint arXiv:2310.00835*.

Wei, Shaohang, Wei Li, Feifan Song, Wen Luo, Tianyi Zhuang, Haochen Tan, Zhijiang Guo, and Houfeng Wang (2025). "TIME: A Multi-Level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios". In: *arXiv preprint arXiv:2505.12891*.

WHO, World Health Organization (2001). "ICF: International Classification of Functioning, Disability and Health". In.

WHO, World Health Organization (2022). "ICD: International Classification of Diseases". In: 11. URL: https://icd.who.int/.

Wilson, Patricia M., Sally Kendall, and Fiona Brooks (2006). "Nurses' Responses to Expert Patients: The Rhetoric and Reality of Self-Management in Long-Term Conditions: A Grounded Theory Study". In: *International Journal of Nursing Studies* 43.7, pp. 803–818.

Wolf, Maryanne, Mirit Barzillai, and John Dunne (2009). "The Importance of Deep Reading". In: *Challenging the Whole Child: Reflections on Best Practices in Learning, Teaching, and Leadership* 130.21.

Woods, Angela (2014). "Beyond the Wounded Storyteller: Rethinking Narrativity, Illness and Embodied Self-Experience". In: *Health, Illness and Disease: Philosophical essays*, pp. 128–113.

Woods, Susan S., Erin Schwartz, Anais Tuepker, Nancy A. Press, Kim M. Nazi, Carolyn L. Turvey, and W. Paul Nichol (2013). "Patient Experiences with Full Electronic Access to Health Records and Clinical Notes Through the My HealtheVet Personal Health Record Pilot: Qualitative Study". In: *Journal of medical Internet research* 15.3, e2356.

Wu, Zhen, Ritam Dutt, Luke M. Breitfeller, Armineh Nourbakhsh, Siddarth Parekh, and Carolyn Rosé (2025). "$R^2 - CoD$: Understanding Text-Graph Complementarity in Relational Reasoning via Knowledge Co-Distillation". In: *International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics*.

Xu, Fangzhi, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria (2025). "Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond". In: *IEEE Transactions on Knowledge and Data Engineering* 37.4, pp. 1620–1634. DOI: 10.1109/TKDE.2025.3536008.

Yao, Hao-Ren, Luke Breitfeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rosé (2024). "Distilling Multi-Scale Knowledge for Event Temporal Relation Extraction". In: *In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2971–2980.

Yuan, Weizhe and Pengfei Liu (2022). "reStructured Pre-training". In: *arXiv preprint arXiv:2206.11147*.

Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. (2020). "Big Bird: Transformers for Longer Sequences". In: *Advances in Neural Information Processing Systems* 33, pp. 17283–17297.

Zhang, Ying, Tao Xiang, Timothy M Hospedales, and Huchuan Lu (2018). "Deep Mutual Learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328.

Zhao, Wenbin, Yuhang Zhang, Di Wu, Feng Wu, and Neha Jain (2025). "Hypergraph Convolutional Networks with Multi-Ordering Relations for Cross-Document Event Coreference Resolution". In: *Information Fusion*. Vol. 115.

Zhou, Jie, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou (2022). "RSGT: Relational Structure Guided Temporal Relation Extraction". In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2001–2010.

Zhou, Li, Genevieve B. Melton, Simon Parsons, and George Hripcsak (2006). "A Temporal Constraint Structure for Extracting Temporal Information from Clinical Narrative." In: *Journal of Biomedical Informatics*. Vol. 39, pp. 424–439.

Zwaan, Rolf A. (1996). "Processing Narrative Time Shifts". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Vol. 22, p. 1196.