Subjectivity in the Annotation of Bridging Anaphora

Lauren Levine and Amir Zeldes
Georgetown University
Department of Linguistics
{lel76, amir.zeldes}@georgetown.edu

Abstract

Bridging refers to the associative relationship between inferable entities in a discourse and the antecedents which allow us to understand them, such as understanding what "the door" means with respect to an aforementioned "house". As identifying associative relations between entities is an inherently subjective task, it is difficult to achieve consistent agreement in the annotation of bridging anaphora and their antecedents. In this paper, we explore the subjectivity involved in the annotation of bridging instances at three levels: anaphor recognition, antecedent resolution, and bridging subtype selection. To do this, we conduct an annotation pilot on the test set of the existing GUM corpus, and propose a newly developed classification system for bridging subtypes, which we compare to previously proposed schemes. Our results suggest that some previous resources are likely to be severely under-annotated. We also find that while agreement on the bridging subtype category was moderate, annotator overlap for exhaustively identifying instances of bridging is low, and that many disagreements resulted from subjective understanding of the entities involved.

1 Introduction

Bridging is an anaphoric phenomenon where a newly introduced discourse entity is dependent on an associated, non-identical antecedent entity for interpretation. The term “bridging” refers to a discourse participant’s construction of an implicature from the entity they are currently processing back to an antecedent entity (Clark, 1975). This associative relation can be triggered by a broad variety of linguistic mechanisms, including lexical part-whole relations (*a house - the door*) and implicit arguments (*a murder - the victim*). Since the phenomenon was first commented on by Clark (1975), it has received a variety of theoretical treatments, including Prince (1981)’s closely related notion of *Inferrables* which centers information status as the key component in identifying anaphoric bridging relations. Such theoretical divides have resulted in a number of different annotation formalisms varying in their definitions of bridging, as well as in their delineations of sub-varieties of bridging (Kobayashi and Ng, 2020). While there has recently been some effort to harmonize bridging annotations across different corpora (Levine and Zeldes, 2024), the current landscape of bridging resources remains heterogeneous. The lack of consistency in and across bridging resources largely stems from their differing definitions for bridging, as well as the subjective annotator judgments that go into identifying instances of bridging.

In this paper, we explore subjectivity in the annotation of bridging anaphora in order to understand how to account for that subjectivity and create more consistent annotations in future efforts. We examine three stages in the annotation process where annotators must make subjective judgments: (1) recognition of the bridging anaphor, (2) resolving back to its associated antecedent, and (3) identifying the subtype category of the bridging pair. To this end, we conduct an annotation pilot on the test set of an existing English corpus, GUM (v10) (Zeldes, 2017). While the GUM corpus includes bridging annotations, the annotation guidelines are underspecified and do not include bridging subtype annotations. This annotation pilot is a preliminary phase in the development a new bridging resource, GUMBridge. For this effort, we develop a new classification system for bridging subtypes organized under 3 relation types: COMPARISON relations, ENTITY relations, and SET relations, as well as an additional OTHER category. We also create annotation guidelines for how to identify instances of bridging anaphor-antecedent pairs and how to classify them into subtypes.

Analyzing the results of this pilot, we find on the one hand that we are able to identify substantially

Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025), pages 48–59
July 31, 2025 ©2025 Association for Computational Linguistics

more and denser attestation of bridging than suggested by several previous resources. In terms of subjectivity, we find moderate agreement for the selection of the bridging subtype category and for the selection of an antecedent for a given anaphor. However, the annotator overlap in the recognition of bridging anaphora is considerably lower, despite mostly plausible precision. We conduct a qualitative evaluation of the annotations from the pilot, and we find that subjectivity plays a role in each of the three annotator judgment stages listed above, especially for recall. We explore this role for each stage, and then give recommendations on how to structure the annotation of bridging anaphora in order to account for subjectivity in annotator judgment.

2 Background

As mentioned above, there are a number of different annotation formalisms for bridging, all with somewhat different definitions of bridging as a phenomenon. In English, the evaluation of bridging resolution systems (systems which aim to automatically identify bridging anaphora and resolve back to their associative antecedents) is commonly conducted using the following three corpora: ISNotes (Markert et al., 2012), BASHI (Rösiger, 2018), and ARRAU RST (Poesio and Artstein, 2008; Uryupina et al., 2019). While ARRAU RST annotates bridging instances by identifying mention pairs that establish cohesion in text and then classifies then via a set of predefined semantic relations, ISNotes and BASHI annotate bridging anaphora based on the information status of entities, considering bridging to be a sub-variety of mediated information.

The information status (IS) of an entity refers to the extent to which the entity is accessible to the reader/hearer of a discourse (Nissim et al., 2004). Generally speaking, "New" information is unrecognized by the reader/hearer, while "Given" information is recognized. "Given" entities may be recognized by the reader/hearer for various reasons: the entity may have been previously introduced in the discourse (coreference), the entity may be accessible via generics/world knowledge, or, in the case of bridging, the referent of the entity may be inferred from a previous entity in the discourse. Instances of bridging and generics/world knowledge are both considered "Accessible" in that they are recognized by the reader/hearer when they are first introduced to the discourse, but only instances of bridging depend on an associative antecedent for comprehension.

	Tokens	Bridging Instances	Bridging per 1k Tokens
ARRAU RST	229k	3.7k	16.5
ISNotes	40k	663	16.6
BASHI	58k	459	7.9
GUM (v10; full)	228k	1.9k	8.3
GUM (v10; test only)	26k	222	8.5
GUMBridge (v0.1)	26k	401	15.4

Table 1: Frequency of bridging instances several English bridging resources.

There are also a number of other existing bridging resources: in English, GUM, SciCorp (Roesiger, 2016), corefpro (Grishina, 2016), RED (Richer Event Descriptions, O'Gorman et al. 2016); as well as in other languages: GRAIN (Schweitzer et al., 2018) and DIRNDL (Eckart et al., 2012) in German, PDT (Nedoluzhko et al., 2009) in Czech, and PCC (Ogrodniczuk and Zawisławska, 2016) in Polish, to name a few. There have additional been efforts in areas closely related to bridging, such as Recasens et al. (2010), which puts forward a typology for classifying near-identity relations (NIDENT) for coreference, and Modjeska (2004)'s work on other-anaphora, which we now consider a subtype of bridging. We provide background on ISNotes, BASHI, and ARRAU RST, as they are commonly used in bridging resolution evaluation (Yu et al., 2022; Kobayashi et al., 2023), and they illustrate diverging perspectives on identifying bridging instances. Table 1 shows comparative statistics for these three resources, the original GUM bridging annotations, and the bridging annotations produced in the GUMBridge annotation pilot described in this paper.

ISNotes is a corpus of 50 Wall Street Journal (WSJ) documents from the OntoNotes corpus (Weischedel et al., 2011) annotated for fine-grained information status. ISNotes distinguishes three main categories of IS: `New`, `Old`, and `Mediated`. `Old` information is that which known to the hearer and/or has been refereed to previously, while `New` information is introduced for the first time. `Mediated` information has not been introduced before, but is not independently comprehensible, requiring either an inference from a previous mention or from general/real-world knowledge. Within the `Mediated` category, there are six subcategories, including `bridging`. The corpus contains 663 instances of bridging in the mediated/bridging category, and there are an additional 253 instances of comparative anaphora in the mediated/comparison category, which is considered a variety of bridging (~16.6 bridging instances per 1k tokens). Markert et al. (2012) report Cohen's κ for annotator pairs, ranging ~0.6-0.7 for mediated/bridging, and ~0.8 for mediated/comparison. They note that the agreement for mediated/bridging is more annotator dependent relative to the other IS categories.

The BASHI corpus is also annotated on top of 50 WSJ documents from the OntoNotes corpus, and it includes a total of 459 bridging pairs (~7.9 bridging instances per 1k tokens). Rösiger (2018) introduces the contrast between referential bridging and lexical bridging, where referential bridging is a properly anaphoric relation (antecedent is required for the interpretation of the anaphor) and lexical bridging is a non-anaphoric semantic relation between two entities. The corpus specifically contains annotations only for referential bridging, not lexical bridging. The bridging instances in BASHI have the subtypes definite, indefinite, and comparative anaphora. Annotator agreement is reported for these categories individually and together. The joint agreement for identifying bridging pairs is 59.3%, with a higher rate for comparative anaphora at 71.4% and lower agreement for definite at 63.8% and indefinite at 42.3%.

ARRAU is a multi-genre corpus covering a variety of anaphoric phenomena, composed of 4 subcorpora, each with its own annotation specifications. ARRAU RST is the largest sub-corpus, and also the one most used in evaluation for bridging resolution. It is composed of WSJ news data, and it includes 3,777 bridging annotations (~16.5 bridging instances per 1k tokens). ARRAU's bridging annotation connects related mentions which establish "entity coherence" via non-identity relations, but as this casts a very broad scope, annotation is limited to a fixed set of semantic relations. The corpus uses an inventory of 9 bridging subtypes for annotation: possession, element-set, subset-set, anaphora marked with 'other', along with accompanying inverse relations of the previous, and an additional under-specified relation. The annotation schema and guidelines for bridging in ARRAU were extended from the GNOME project (Poesio, 2004). Coders in the GNOME project displayed high agreement (95.2%) in the choice of bridging subtype labels from its fixed set of relations, but low recall (22%) in unanimously identifying instances of bridging.

Limiting annotation to a predefined set of relations restricts the scope of bridging as a phenomenon, but also aims to increase consistency in the annotation. However, as has been noted in Rösiger (2018), annotating from predefined relations can also introduce false positives, in the case that an instance of a semantic relation is not actually a case of associative anaphoric reference that would constitute referential bridging. For instance, the case of *Europe - Spain* displays a meronomy relation, but it is not anaphoric because *Spain* can be interpreted without reference to *Europe*. Annotating from an information status informed perspective aims to avoid such false positives, providing a more concrete linguistic criteria for identifying instances of bridging when compared to the notion of "entity coherence", and eliminating the need to only annotate a predefined set of relations for scoping reasons. However, this information status based approach also greatly widens the scope of what should be considered bridging, which in turn increases the influence of subjective judgment by annotators. As such, in order to forward an information status informed annotation perspective, we must develop means of dealing with additional subjectivity it produces.

As we can see in Table 1, there has been considerable variation in the frequency of bridging annotations in previous resources, with ARRAU RST (counting both lexical and referential bridging) and ISNotes identifying bridging instances with approximately twice the rate per 1k tokens as the annotations in BASHI and GUM v10. This suggests that some previous bridging resources, such as BASHI and GUM, have likely been under-annotated for bridging instances and prompts a need for the reexamination of bridging annotation procedures.

3 Annotation Pilot

The analysis on subjectivity in the annotation of bridging instances in this paper is conducted using the results of an annotation pilot for the creation of a new bridging resource called GUMBridge. Built on top of GUM, an existing multi-genre corpus of English, GUMBridge aims to unite aspects of currently existing formalisms: using an information status-informed view of identifying bridging instances (as in ISNotes and BASHI), followed by subtype categorization using a taxonomy of semantic relations (as in ARRAU). Additionally, GUMBridge aims to add genre diversity to the core English bridging resources, as ISNotes, BASHI, and ARRAU RST are all composed of WSJ news data from more than 30 years ago, offering little to analyze in terms on genre diversity. While the development of this resource is still underway, an adjudicated version of the bridging annotations for the GUMBridge test set (version 0.1) is released with this paper[1]. The details of this adjudication process are described in Section 3.5. The guidelines for identifying instances of bridging (v0.1) are described in Section 3.1, and the classification system for bridging subtypes (v0.1) is described in Section 3.2.

3.1 Identifying Bridging Instances

In the GUMBridge annotation effort, we adopt an information status-informed perspective on identifying instances of bridging anaphora. As stated in Section 2, the information status of an entity refers to the extent to which an entity is accessible to the reader/hearer of a discourse upon its introduction. We say that an entity is "Accessible" if it has not been mentioned before but its reference is inferable for a reader/header. Bridging occurs when the first mention of an entity is "Accessible" via an inference from a previous, non-identical entity in the discourse. In contrast with entities which are accessible due to being generic, or being part of world knowledge or the discourse situation, the bridging anaphor is not accessible by itself, but dependent on the previous entity for interpretation. Annotators are provided with an overview of this definition of bridging and accessibility and are instructed to consider the following when deciding whether a particular entity is a bridging anaphor:

1. Do you judge this entity to be to some degree accessible in the discourse?

2. Does that accessibility rely on the understanding of a previous entity in the discourse? If so, identify that previous entity's most recent mention.

If the entity passes the above criteria, it is a bridging anaphor and the previous entity is its associative antecedent. Once identified, a bridging pair can then be assigned a subtype category as described in the following section.

3.2 Classification of Bridging Subtypes

In order to categorize the varieties of bridging present in GUMBridge, we create a new classification system for bridging subtypes. The classification system is composed of 11 categories, 10 of which are organized under 3 relation types: COMPARISON relations, ENTITY relations, and SET relations, and an additional OTHER category. The bridging subtype classification system developed for GUMBridge (v0.1) is shown in Figure 1. A brief description of each of the bridging subtypes follows below. A brief comparison to the bridging subtypes of ARRAU is included in Appendix C.

COMPARISON-RELATIVE The anaphor is preceded by a comparative marker (other, another, same, more, etc.), ordinal (second, third, etc.), or comparative adjective (larger, smaller, etc.), which implies a comparison to the antecedent (or vice versa).

(1) Several women walked into the room. **Other women** soon followed.

COMPARISON-TIME The anaphor refers to a specific time/time frame which is understandable with reference to the time/time frame expressed by the antecedent (or vice versa).

(2) I went shopping Wednesday, March 3rd. I will go again **the following Wednesday**.

COMPARISON-SENSE The type of the anaphor is omitted but inferable via comparison to the antecedent (or vice versa).

(3) I've been to the Chinese restaurant. I want to go to **the Italian one**.

ENTITY-ASSOCIATIVE The anaphor is an attribute or closely associated entity of the antecedent (or vice versa). This frequently manifests as implicit arguments of a predicate as in example (4), relational nouns as in example (5), and prototypical associations as in example (6):

(4) There was a murder last night. **The victim** has yet to be identified.

(5) There is a child in the park. **The parent** must be nearby.

(6) I went to a wedding last week. **The reception** was really fun.

[1] https://github.com/lauren-lizzy-levine/gumbridge

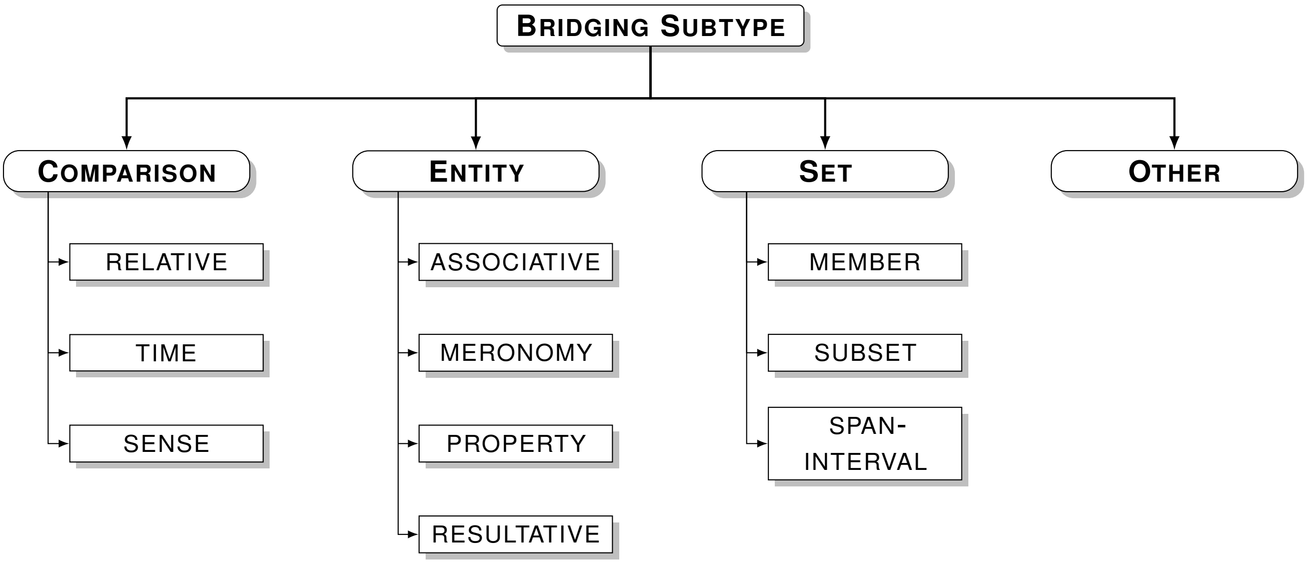

Figure 1: Bridging Subtype Classification in GUMBridge v0.1.

ENTITY-MERONOMY The anaphor is a subunit of the antecedent (or vice versa), i.e., there is some part-whole relation between the anaphor and the antecedent.

(7) I saw a large house by the lake. **The door** was red.

ENTITY-PROPERTY The anaphor is a physical or intangible property of the antecedent (or vice versa). For example: smell, length, style, etc.

(8) I picked up a bouquet of roses. **The scent** was lovely.

ENTITY-RESULTATIVE The anaphor is logically inferable from the antecedent (or vice versa). This is typically the result of a transformative or product producing process, such as cooking.[2]

(9) Though my flour was a strange texture, **the bread** came out perfectly.

SET-MEMBER The anaphor is an element of the antecedent set (or vice versa).

(10) I got several books for my birthday. **The mystery novel** was my favorite.

SET-SUBSET The anaphor is a subset of the antecedent set (or vice versa).

(11) A group of students entered the hall. **The boys** wore neckties with their uniforms.

SET-SPAN-INTERVAL The anaphor is a sub-span of the spatial or temporal antecedent interval (or vice versa).

(12) If you want to meet up on Sunday, I will be free in **the morning**.

OTHER The anaphor and antecedent fit the criteria for identifying a bridging pair, but do not fall into any of the bridging subtypes detailed above. For instance, Ogrodniczuk and Zawisławska (2016) give examples of metareference:

(13) I went to Sensational Cakes yesterday, but I didn't think **the cakes** were very good.

Metareference allows for reference back to a name or label, as in example (13). Such instances are unique and interesting enough to wish not to shoehorn them into another category, but are not common enough to warrant a separate category in the subtype classification.

As stated in Section 3.1, the criterion for identifying instances of bridging is anaphoric, relying on information status and resolution back to an associative antecedent. The subtype labels primarily allow us to understand how the phenomenon manifests in a discourse, and, as such, there is no theoretical reason to limit the number of subtypes that can apply to an instance of bridging to just one. Indeed, there are cases of bridging where multiple subtypes may apply:

(14) Several women walked into the room. **One** left immediately.

[2] This subtype subsumes the TRANSFORMED type proposed by Fang et al. (2022) specifically for recipe outcomes.

(15) I will come to visit this week, as I could not come **the previous week**.

Example (14) shows an instance for which COMPARISON-SENSE and SET-MEMBER both apply, while example (15) show a case where COMPARISON-RELATIVE and COMPARISON-TIME apply. In this annotation pilot, annotators where instructed to select a single bridging subtype, prioritizing certain categories over others if they occurred together. However, in principle, all applicable subtypes could be annotated. In our subsequent efforts to annotate the remaining data in GUM and produce a full version of GUMBridge, we intend to support the annotation of multiple bridging subtypes for a single bridging pair for the entire corpus.

3.3 Annotation Procedure

The GUMBridge annotation pilot was conducted on the test set of the existing GUM (v10) corpus, which consists of 26 documents (~26k tokens) across 16 genres (academic writing, biographies, courtroom transcripts, essays, fiction, how-to guides, interviews, letters, news, online forum discussions, podcasts, political speeches, spontaneous face to face conversations, textbooks, travel guides, and vlogs). The GUM corpus already includes annotations for entity spans, coreference,[3] and information status, i.e., "New", "Given", and "Accessible" (not including accessibility from instances of bridging).

The documents of the test set were double annotated, with one author of this paper acting as Annotator A and various linguistics graduate students acting as Annotator B for different documents in the test set. Each of the 8 annotators acting as Annotator B was assigned between 2 and 4 documents of the test set. The annotation was completed using the GitDox annotation interface (Zhang and Zeldes, 2017). For the existing entity annotations in the document, the annotator was instructed to identify whether the entity is a bridging anaphor, and, if so, create a link between the anaphor and its associative antecedent. The annotator was instructed to also update the IS of the bridging anaphor to "Accessible" and select a bridging subtype annotation for the anaphor. The full annotation guidelines provided to the annotators are included as supplementary materials.

3.4 Agreement Study

In Table 2, we provide agreement numbers for three stages of the bridging annotation process: anaphor recognition, antecedent resolution, and subtype categorization.

	Precision	**Recall**	**F1 Score**
Anaphor Recognition	0.44	0.34	0.38
Anaphor+Antecedent Recognition	0.32	0.25	0.28
	Accuracy		
Antecedent Resolution	0.72		
	Cohen's κ		
Bridging Subtype	0.58		

Table 2: GUMBridge pilot inter-annotator agreement.

For the recognition of bridging pairs (anaphor+antecedent) and recognition of the bridging anaphor alone, we give the PRF of Annotator B relative to Annotator A. We see that the F1 for bridging anaphor recognition is 0.38, and the F1 for bridging pair recognition is only 0.28. As the recognition of bridging pairs is inherently limited by the recognition of the anaphor, we also give the accuracy of Annotator B selecting the antecedent entity when both annotators agree on the bridging anaphor, which is 72% of a total of 133 cases. Finally, for the 96 instances where both annotators agreed on the anaphor and antecedent of a bridging pair, the Cohen's Kappa for the bridging subtype annotation is 0.58, which indicates moderate agreement. These numbers suggest that the key hurdle is in anaphor recognition, though antecedent resolution and subtype labeling are also non-trivial.

In Figure 2, we show a confusion matrix of the bridging subtype labels assigned by Annotator A and Annotator B to the overlapping bridging pairs. We see that the subtypes with the most overlap are the COMPARISON categories and ENTITY-ASSOCIATIVE. And while there are some categories for which the disagreement is spread among a number of categories, we see that the categories of ENTITY-MERONOMY and SET-MEMBER are particularly confusable, which indicates how part-whole and set-member relations can be quite similar. The categories of ENTITY-ASSOCIATIVE and OTHER are also particularly confusable, which

[3] The coreference scheme considers all mentions eligible for bridging, including indefinite anaphors, discourse deixis to non-nominal antecedents and more, see Zeldes (2022) for a detailed discussion.

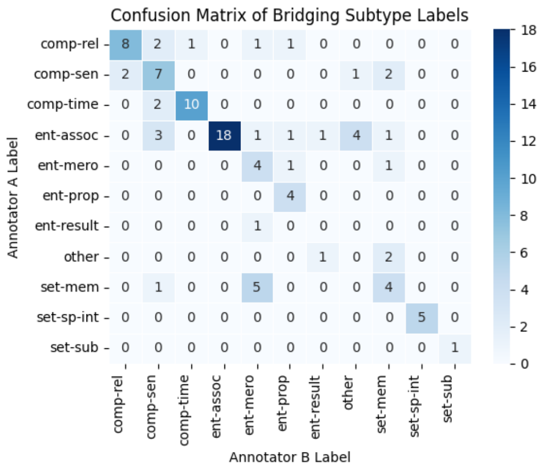

Figure 2: Confusion matrix of bridging subtypes for bridging instances with matching anaphor and antecedent annotations.

speaks to how ENTITY-ASSOCIATIVE may be an overly broad category. Although agreement on bridging subtype annotation is moderate, it is clear that refinement in the guidelines for the categories is still needed. However, as agreement on the identification of bridging instances is substantially lower, recognition of bridging anaphora forms the limiting point in the annotation process.

3.5 Data Adjudication

As shown in the previous section, the results of the annotation pilot had low annotator agreement, necessitating a qualitative analysis of annotations to determine the cause of the disagreements. As a part of this process, the annotations from the pilot were adjudicated to produce a single set of reference bridging annotations for the test set of GUMBridge (v0.1), available with the release of this paper under the Creative Commons Attribution (CC-BY) version 4.0 license. The composition of the GUMBridge test set by bridging subtype after the adjudication is shown in Appendix A. The test set of GUMBridge has a total of 401 bridging annotations, with an average of 15.4 bridging instances per 1k tokens. This is on par with the higher rate of bridging instances per 1k tokens found in IS-Notes and ARRAU RST as shown in in Table 1. While the limited size of the data set annotated in this pilot limits our ability to make observations on genre effects, for completeness, a breakdown of the bridging relation types observed in each genre is included in Appendix B .

Notably, the number of instances in the test set of the GUM (v10) annotations nearly doubles, going from 222 instances of bridging to 401 in GUMBridge test, suggesting a significant improvement in coverage of bridging instances in this new annotation effort. Even though there is less consistency in this annotation effort compared to some of those discussed in Section 2, numbers suggest higher recall, which allows us to capture a greater scope of bridging instances. As bridging is generally a sparse phenomenon, the annotations can be manually reviewed and validated in the adjudication process even if initial agreement is low. As such, we believe it is preferable to favor a high recall method of annotation and eliminate false positives upon review, rather than risk many interesting cases that will remain unidentified.

Completely Matching	61
Different Subtype	35
Different Antecedent	37
Annotator B Only	172
Annotator A Only	257
Total	562

Table 3: Counts of annotator agreement/disagreement types in GUMBridge pilot annotations.

The adjudication process involved comparing all of the diverging judgments from Annotator A and Annotator B at the level of anaphor, antecedent, and subtype. Table 3 shows the proportion of such disagreements in the pilot annotations. Of the 172 instances that Annotator B labeled as bridging which Annotator A initially did not label as bridging at all, upon reevaluation, it was concluded that 64 (37%) could reasonably be considered a form of bridging. Many of these judgments relied on subjective understanding of the discourse entities involved. In the following section, we provide an analysis of the impact of subjectivity in this annotation pilot and how it may be better handled in the future.

4 Subjectivity in Bridging Annotation

Previous work on subjectivity in the development of linguistic data has heavily featured areas where annotator judgments can be highly variable, such as hate speech detection and sentiment analysis (e.g., Waseem (2016); Kenyon-Dean et al. (2018)), though attention has also been given to tasks which seem more objective, such as part of speech annotation (e.g., Plank et al. (2014)). Several works discuss the paradigms for and implications of including subjective judgments in annotation efforts, rather than trying to eliminate all ambiguity (Ovesdotter Alm, 2011; Röttger et al., 2022). Ultimately, the appropriate approach depends on the linguistic task at hand and what the researchers are hoping to achieve with the annotation effort.

Although detailed guidelines are provided to annotators in this paper's annotation pilot, subjective judgment is still an inherent part of the annotation of bridging instances, as annotators are making decisions based off their understanding of the implicit relationships that exist between entities in a discourse. As previously noted, there are three decision points in the annotating of bridging instances that can introduce subjective judgment: (1) recognition of the bridging anaphor, (2) identifying the corresponding associative antecedent, and (3) selecting the bridging subtype category of the pair. The sections below give examples to illustrate the unique considerations regarding subjectivity that are present at each of these annotation stages.

4.1 Subtype Categorization

Selecting a bridging subtype category relies on understanding the relationship between the anaphor and the antecedent in a bridging pair. The exact nature of the relationship between two entities is dependent on the annotator's subjective conception of the two entities. It is possible that a lack of familiarity with related entities may cause annotation errors:

(16) the cuttings → **the first pad**

In example (16), "the cuttings" refer to cactus cuttings, each of which is a whole pad. Without this particular knowledge, it would be reasonable for an annotator to assume that a pad is a portion of a cutting or that a cutting is a portion of a pad.

There may be additional uncertainty in interpreting an entity based on the context of the discourse:

(17) peppermint plants → **the mint**

In the discourse context of example (17), it is unclear whether "the mint" is referring back to a specific part of the peppermint plant (e.g. the leaves), or whether it is an instance of synecdoche, referring to the plant as a whole.

There are also instances where multiple subtypes are possible in the context of the discourse:

(18) some basil → **seed**

In the discourse context of example (18), a question is being posed whether "some basil" can be grown from "seed". As such, it is reasonable to say that the basil comes from the seed in which case the subtype would be ENTITY-RESULTATIVE. However, it is also reasonable to say that seed is a part of the basil plant, in which case the subtype would be ENTITY-MERONOMY. In such cases, it is necessary to have a priority hierarchy for deciding which bridging subtype category should be assigned, or we must allow for multiple subtype annotations. In future work, we intend to support the annotation of multiple bridging subtypes for the entire GUMBridge corpus.

4.2 Antecedent Selection

When an annotator is selecting the associative antecedent of a bridging anaphor, there are also opportunities for subjective judgments to be made. In some cases, it is possible that multiple preceding entities could be reasonable candidates for a bridging antecedent:

(19) your mouth → **other body parts...**
teeth → **other body parts...**

The example (19) refers to a case where a dental cast is being made and the narrator wonders what other body parts can be given the same treatment. It is not clear whether "the other body parts" are more appropriately in contrast with the "mouth" or "teeth", or even both, if we accept both teeth and mouths as body parts.

There is also the possibility for disagreement on the denotation of the anaphor:

(20) the bridge → **the edge**
the upper levels → **the edge**

In example (20), the narrator considers looking over "the edge", and it is unclear whether it is the edge of a particular bridge, or if it is the edge of some general upper level. In such cases, it may be beneficial to impose an easy to execute heuristic, such as selecting the option nearer to the bridging anaphor, assuming we are aiming for a single reference decision. Note that this is different from cases in which multiple labels apply, since the two interpretations, while both possible, are mutually exclusive.

4.3 Anaphor Identification

When identifying a bridging anaphor, annotators must make subjective judgments on whether an entity is accessible due to world knowledge (and hence not bridging) or whether the accessibility can be attributed to an antecedent entity. For instance, one annotator had “Leucippus and Democritus” bridge from “ancient Greek philosophers”, but not “Aristotle” who is more widely known. This illustrates how an annotator’s world knowledge may influence what they consider to be “Accessible” in a manner that is undesirable as it will lead to inconsistencies among annotators. We recommend that concrete criteria for generic/world knowledge accessibility should be tied to a knowledge base, such as Wikipedia, rather than left up to individual annotator judgment. For named entities, this type of linking or Wikification is already available for GUM (Lin and Zeldes, 2021) and will be integrated in future annotation efforts.

5 Conclusion

In this paper, we examine the influence of subjectivity in annotator judgment on the various stages of annotating instances of bridging. We make this examination using the resulting annotations from a pilot to create a new resource for bridging annotations, GUMBridge. We also release an adjudicated version of the bridging annotations for the preliminary test set of GUMBridge (v0.1). In subsequent work, we plan to refine the guidelines and annotation procedure used in this pilot, which we will then use to annotate the remainder of the GUM corpus (dev and train) to produce a full version of GUMBridge, as well as extending our annotations to GUM’s out-of-domain challenge test set, GENTLE (GEnre Tests for Linguistic Evaluation, Aoyama et al. 2023). As the time and effort required to manually annotate bridging limits the scalability of the annotation process, we will also investigate incorporating semi-automated methods, such as combining LLMs or other systems for bridging resolution with human correction in order to improve the efficiency of the process.

In our development of GUMBridge test (v0.1), we found that annotators’ agreement on selecting the subtype of a bridging pair was moderate, but that it was more difficult to get the annotators to align on the identification of bridging anaphora. This indicates that recognition of bridging anaphora is the stage in the annotation process that is most vulnerable to the subjective judgment of annotators, and that should be given the most consideration when trying to account for annotator subjectivity. While some subjectivity arises from the inherent ambiguity of language in context, other aspects of subjectivity can be accounted for by providing guidelines on how to decide on preferable judgments when multiple options are available.

Limitations

The analysis presented in this paper on subjectivity in the annotation of bridging anaphora is based on a pilot annotation study for a new resource that is still in development. This limits the amount of data available for analysis to a test set of 26k tokens. The reliability of the annotation schema is also a limitation, as the results of the annotation pilot showed agreement on identification of bridging anaphora to be undesirably low, and the annotation schema/instructions will need to undergo revision in future work.

References

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. *A Discourse Information Radio News Database for Linguistic Analysis*, pages 65–76. Springer Berlin Heidelberg, Berlin, Heidelberg.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Yulia Grishina. 2016. Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 7–15, San Diego, California. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It‘s complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers), pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2023. PairSpanBERT: An enhanced language model for bridging resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6931–6946, Toronto, Canada. Association for Computational Linguistics.

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lauren Levine and Amir Zeldes. 2024. Unifying the scope of bridging anaphora types in English: Bridging annotations in ARRAU and GUM. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 41–51, Miami. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for wikification in 12 genres. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Natalia N Modjeska. 2004. Resolving other-anaphora.

Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 108–111, Suntec, Singapore. Association for Computational Linguistics.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Maciej Ogrodniczuk and Magdalena Zawisławska. 2016. Bridging relations in Polish: Adaptation of existing typologies. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 16–22, San Diego, California. Association for Computational Linguistics.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, pages 223–255.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ina Roesiger. 2016. SciCorp: A corpus of English scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1743–1749, Portorož, Slovenia. European Language Resources Association (ELRA).

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. German radio interviews: The GRAIN release of the SFB732 silver standard collection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Joseba Rodríguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26:95 – 128.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Shuo Zhang and Amir Zeldes. 2017. GitDOX: A linked version controlled online XML editor for manuscript transcription. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2017)*, pages 619–623.

A Subtypes in GUMBridge Test

Table 4 shows the counts of the bridging subtypes in the adjudicated version of GUMBridge test v0.1.

B Subtypes by Genre in GUMBridge Test

Figure 3 shows the number of bridging instances per 1k tokens of each bridging relation type (COMPARISON, SET, ENTITY, and OTHER) in each of the 16 genres in GUMBridge test (v0.1).

COMPARISON	
RELATIVE	59
TIME	27
SENSE	45
Subtotal	131
ENTITY	
ASSOCIATIVE	124
MERONOMY	37
PROPERTY	9
RESULTATIVE	21
Subtotal	191
SET	
MEMBER	31
SUBSET	14
SPAN-INTERVAL	18
Subtotal	63
OTHER	16
Total	401

Table 4: Counts of bridging subtypes in adjudicated GUMBridge data.

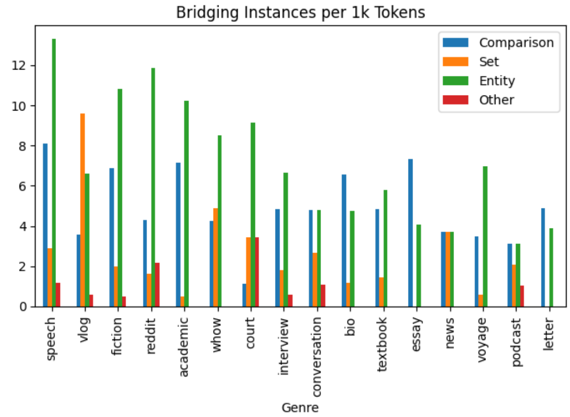

Figure 3: Counts of bridging relation types by genre in adjudicated GUMBridge data.

C Comparison with ARRAU Bridging Subtypes

In order to allow for better comparison between the resources of GUMBridge and ARRAU, we include a brief comparison of how ARRAU's bridging subtypes[4] map onto the proposed schema for GUMBridge:

[4] As the GUMBridge schema does not differentiate the relative roles of the anaphor and antecedent in the subtype relation, ARRAU's inverse subtypes map the same as their regular subtypes.

possession → Part-of relations that will mostly fall under ENTITY-MERONOMY or ENTITY-PROPERTY.

element-set → Maps to SET-MEMBER.

subset-set → Maps to SET-SUBSET.

'other' anaphora → Maps to COMPARISON-RELATIVE, which encompasses additional comparative markers not covered in ARRAU, including ordinals and comparative adjectives.

under-specified → ENTITY-ASSOCIATIVE unless one of the other ENTITY subtypes is a better fit based on the context. However, sense anaphora (green shirt → **red one**) should be mapped to COMPARATIVE-SENSE.