What's in a Bridge?: A Descriptive, Multi-Genre Analysis of the GUMBridge Corpus for Varieties of Bridging Anaphora

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

Abstract

In this paper, we present a descriptive corpus analysis of bridging anaphora across 16 genres of English, leveraging the multi-genre GUMBridge corpus for varieties of bridging anaphora. We begin our investigation by examining the distribution of bridging instances by sub-varieties and across genres, finding that spoken genres have less bridging instances than written ones. We then investigate the linguistic environments of bridging anaphora and their corresponding associative antecedents in the underlying data of the corpus, examining both categorical features (entity type, part of speech, syntactic dependency relations) and numeric features (mention length, cluster size, salience, and distance between the bridging anaphor and antecedent). We find bridging anaphora have a tendency to be shorter and are more often definite, and bridging antecedents show a tendency to be more salient than other entities. Finally, we analyze how several of the numeric features of bridging environments vary by genre, finding consistent patterns across genres for observed trends in the environments of bridging anaphora and antecedents.

1 Introduction

"Bridging" refers to an anaphoric phenomenon where the referent of a newly introduced entity (the bridging anaphor) is inferable specifically due to its relationship with a previously introduced entity in the discourse (the associative antecedent). Consider the following example:

(1) There is a house. **The door** is red[1].

In the example above, **the door** is understood specifically to be the door of the aforementioned house. The referent of **the door** cannot be resolved without reference to a house. A "bridge" is constructed from the entity that is currently being processed back to an antecedent entity, resulting in an associative relation between the two entities (Clark, 1975). This associative relation can manifest in different manners in a discourse, including lexical part-whole relations (e.g, a house → **the door**), implicit arguments (e.g, a murder → **the victim**), and sense anaphora (e.g, a red ball → **a blue one**).

Bridging 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 bridging relations. There are a number of English bridging resources which take an information status informed approach to identifying anaphoric bridging, referred to as *referential bridging* (Rösiger, 2018). Until recently, prominent corpora for referential bridging, such as ISNotes (Markert et al., 2012) and BASHI (Rösiger, 2018), have been too small (916 and 459 bridging instances respectively) and lacking in genre diversity (both Wall Street Journal news data) to be suitable for meaningful, descriptive corpus analysis of the phenomenon.[2]

However, the recent GUMBridge corpus (Levine and Zeldes, 2026) contains 5.7k instances of bridging across 24 genres of English, providing new opportunities for a rich, genre diverse analysis of bridging anaphora in English. We aim to provide that analysis in this paper. We first examine the distribution of bridging instances in the GUMBridge corpus across genres and bridging sub-varieties (Section 4). We then analyze the linguistic environments of bridging anaphora and their associative antecedents to investigate when/where bridging occurs (Sections 5.1, 5.2, and 5.3). Finally, we conclude our investigation by analyzing how the linguistic features of bridging environments vary by genre and modality (Section 5.4).

[1] Bridging anaphora are marked in bold face, and their associative antecedents are underlined.

[2] A number of resources also exist for other languages, such as German (Björkelund et al., 2014) and Czech (Nedoluzhko et al., 2009), which we do not discuss for space reasons.

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 40–52

2 Background

Clark (1975) offers the first theoretical account of bridging as a phenomenon, covering a broad range of discourse inference. Subsequently, there has been a variety of theoretical accounts of bridging which have provided different perspectives on the phenomenon; for instance, Hawkins (1978) discusses associative entity relations (i.e., bridging) as a means of licensing of the definite article *the* with newly introduced entities, Asher and Lascarides (1998) formalizes bridging in the SDRT framework (Asher, 1993), and Baumann and Riester (2012) discusses bridging in the context of a two level (referential and lexical) information status designation.

As mentioned above, in addition to theoretical discussions connecting bridging and information status (Baumann and Riester, 2012; Prince, 1981), prominent English corpora for bridging anaphora rely on information status based definitions of bridging. The information status of an entity refers to the extent to which the entity is accessible to the reader/hearer of a discourse (Nissim et al., 2004). An entity is either "New" information, where the entity is not yet known to the participant, or "Given" information, where the entity is recognized by the participant. However, when an entity is newly introduced but the referent is still inferable, it is said to be "Accessible". In the case of a bridging anaphor, the entity is Accessible specifically due to its relation to a previous entity in the discourse, making it an anaphoric phenomenon.

However, not all English bridging resources use an information status based definition of bridging. The ARRAU corpus annotates related mentions that establish entity coherence through non-identity relation as bridging (Poesio and Artstein, 2008; Uryupina et al., 2019). Rösiger et al. 2018 introduces the distinction between referential and lexical bridging as a means of describing the differences in the bridging definitions used by these corpora. Referential bridging refers to truly anaphoric instance of bridging where the bridging anaphor requires an antecedent to be interpretable, as in (2):

(2) She likes the house because **the windows** are large.

Lexical bridging refers to lexical semantic relations between pairs of entities, such as part-whole or set-member relations, which may or may not be anaphoric, as in (3) where the antecedent is not strictly necessary for interpretation:

GUMBridge	Tokens	Bridging Instances	Bridging per 1k Tokens
Train	213k	4k	18.9
Dev	30k	732	24.5
Test	30k	562	18.6
Total	273k	5.3k	19.5

Table 1: Distribution of bridging instances across GUMBridge main corpus partitions.

(3) I went to the United States last month. My first stop was **Washington, DC**.

The GUMBridge corpus is entirely composed of instances of referential bridging.

3 Data

The GUMBridge corpus is recent effort to annotate sub-varieties of bridging anaphora in English (Levine and Zeldes, 2026). It is built on top of the multi-genre GUM v12 corpus (Zeldes, 2017), which contains a variety of linguistic annotations, including part of speech, syntactic Universal Dependencies relations (UD, de Marneffe et al. 2021), discourse parses, and various entity annotations. The main corpus (train, dev, and test partitions) contains 5.3k instances of bridging across 16 genres and 273k tokens of English (see Table 1 for the partition sizes). While there is also an extended test set (test2; ~18k tokens) with an additional 8 genres (GENTLE; Aoyama et al. 2023), we do not include it in this analysis due to the limited data size of the individual genres. Further mentions of the GUMBridge corpus as a whole will refer to the data in the collective partitions of train, dev, and test. These partitions contain the following 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.

GUMBridge uses an information status based, anaphoric definition of bridging known as referential bridging (see Section 2). It requires that a bridging anaphor has the information status "Accessible", meaning that its referent is inferable upon first mention. Additionally, that accessibility must be specifically due to a relationship to a previous, non-identical entity. This contrasts with the phenomenon of entity coreference (Zeldes, 2022),

Multi-subtype Count	Occurrences
1 subtype	4,547
2 subtypes	733
3 subtypes	39
4 subtypes	2
Total Instances	5,321

Table 2: Bridging multi-subtypes in GUMBridge.

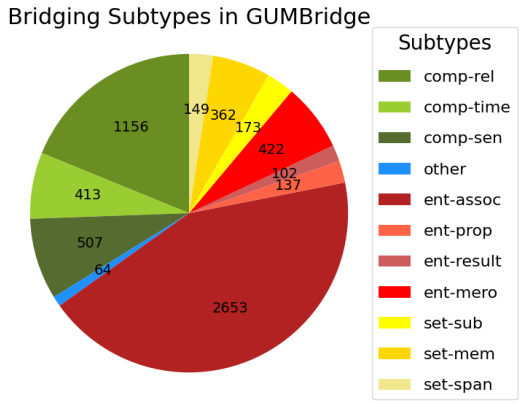

Figure 1: Bridging subtype proportions in GUMBridge. Raw counts for each subtype annotation are also shown.

where there is an identity relation to a previous entity, and the information status of the second entity is "Given", i.e., already known.

GUMBridge also uses a categorization schema for bridging sub-varieties with 3 main categories of bridging relations:

COMPARISON Relations The anaphor is preceded by a descriptor which implies a comparison to the antecedent (or vice versa) (e.g., a tree → **a taller tree**).

ENTITY Relations The anaphor is an attribute or associated entity of the antecedent (or vice versa) (e.g., a cafeteria → **the food**).

SET Relations There is a set/subset or membership relation between the bridging anaphor and antecedent. (e.g., several flowers → **the rose**).

Within these three main categories, GUMBridge distinguishes 10 sub-varieties, and there is an additional OTHER category for a total of 11 sub-varieties (see Appendix A for details). These subtype labels are a means of understanding how the phenomenon of bridging manifests in a discourse. As the criteria for identifying bridging is anaphoric, there is no theoretical need to limit the number of

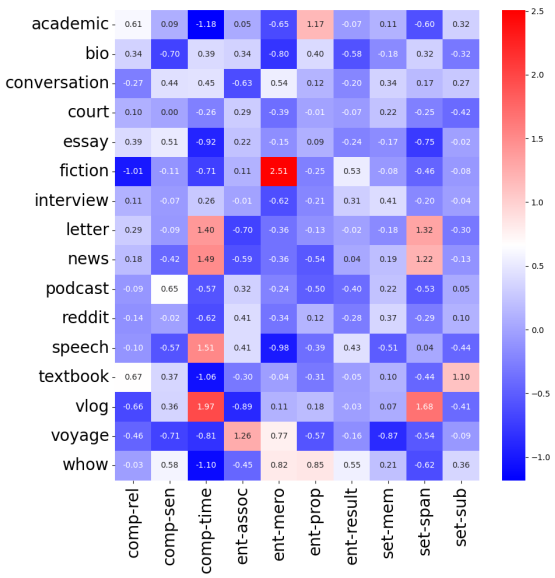

Figure 2: χ^2 residuals of bridging subtype occurrences by genre.

subtypes that can apply to an instance of bridging. As such, every instance of bridging in the GUMBridge corpus has a subtype annotation which contains one or more subtype labels. Table 2 shows the count of multi-subtype bridging instances in GUMBridge. In the following sections, we leverage the bridging instance and bridging subtype annotations of GUMBridge along with the various linguistic annotations of the underlying GUM corpus to analyze the distribution and environments of bridging instances in the full GUMBridge data set and across genres.

4 Genre and Subtype Distribution of Bridging Anaphora

Subtypes First, we will consider the overall distribution of bridging subtypes in the GUMBridge corpus. In Figure 1, we show the proportions and raw counts of the subtype label annotations in the GUMBridge corpus. Since multiple subtypes can be applied to a single instance of bridging (see Table 2), we count each subtype label individually. We see that ENTITY relations are dominant, comprising 54.0% of labels, followed by COMPARISON relations with 33.8% of labels, and finally SET relations with 11.2% of labels (the remaining 1.0% of labels are OTHER).

The most common subtypes are ENTITY-ASSOCIATIVE with 43.2% of labels and COMPARISON-RELATIVE with 18.8% of labels. Part of the reason the ENTITY-ASSOCIATIVE

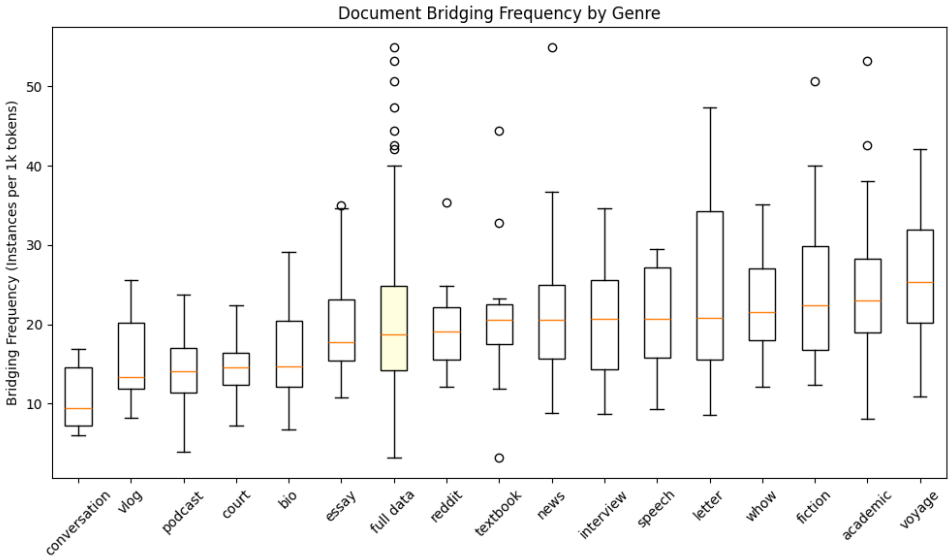

Figure 3: Document frequency of bridging instances per GUMBridge genre.

subtype makes up such a large proportion is that it is a broadly construed category, containing various associative relations, such as relational nouns (e.g., a business → **the customer**), implicit arguments (e.g., a murder → **the victim**), and prototypical/inducible associations (e.g., a wedding → **the reception**). While more insight may be gained by further delineating this subtype, it is still notable that these types of associative entity relations are so pervasive in the annotation of bridging instances. COMPARISON-RELATIVE relations occur when there is an overt comparative marker that must be resolved by an implicit inference. For instance, in the sentence "I like your dog, but I want **a bigger dog**.", "a bigger dog" is understood to mean "a bigger dog than your dog". It makes sense that this is a common subtype, as relative adjectives and other such comparative markers are common in English. The least common subtype is ENTITY-RESULTATIVE with 1.7% of labels. The entity-resultative subtype is narrow in scope, focusing specifically on causal/transformational relations between entities, which are frequently seen in contexts involving product producing processes, such as cooking/baking (e.g., some flour → **the bread**, Fang et al. 2022), but not common in most contexts.

Distribution across genres Generally speaking, bridging is considered to be a relatively rare phenomenon. As we can see in Table 1, bridging instances occur at an overall density of only 19.5 instances per 1k tokens in the GUMBridge corpus. However, this density is not uniform across documents and genres, as we can see in Figure 3, which shows the distribution of per document bridging instances across genres. Across the 16 genres in GUMBridge, the median document bridging frequencies per 1k tokens have a relatively broad range: from *conversation* with 9.5 on the low end up to *voyage* (travel guides) with 25.4 on the high end, while the median document frequency for the full data set is of 18.7. On the lower end of the spectrum, we see a clustering of spoken genres: *conversation*, *vlog*, *podcast*, and *court*, while more structured, written genres such as *academic* and *voyage*, cluster at the top.

Subtypes versus genres In Figure 2 we show a heatmap of χ^2 residuals of the occurrences of bridging subtype labels across genres. We see that ENTITY-MERONOMY instances of bridging occur most prominently in the genre of *fiction* (χ^2 residual +2.51). The genres of *letter*, *news*, and *vlog* show elevated residuals for the subtypes of COMPARISON-TIME and SET-SPAN-INTERVAL. These subtypes frequently pattern together when the anaphor overlaps with the time frame of the antecedent (e.g., this week → **Thursday**), and it makes sense that such time comparisons are common in genres like *letter* and *news*, where explicitly providing a date of writing or publication is common. A number of the more negative residuals are also for COMPARISON-TIME in genres were there are not typically anchoring times at the outset of a document, such as *whow* (how-to guides from Wikihow) and *academic*.

	Precision	Recall	F1-Score
Anaphor			
Random Forest	72.3	60.1	65.6
Random Baseline	51.2	51.3	51.2
Antecedent			
Random Forest	60.9	49.9	54.8
Random Baseline	50.5	48.7	49.6

Table 3: Performance of bridging anaphor and antecedent detection random forest classifiers on GUMBridge test.

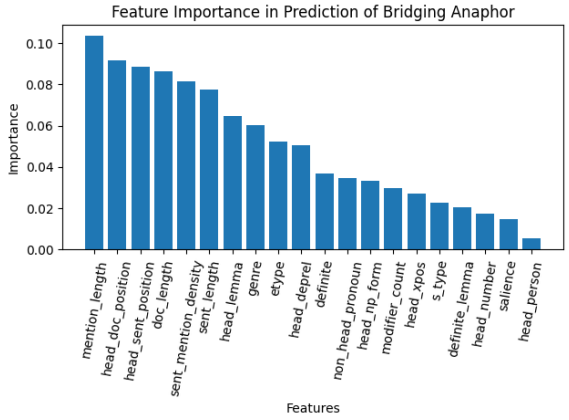

Figure 4: Feature importance for random forest classifier detecting bridging anaphor mentions.

5 Feature Environments of Bridging Anaphora and Associative Antecedents

In this section we model the environments in which bridging anaphora occur using several multifactorial tree ensemble models (Section 5.1), followed by in depth analysis of categorical (Section 5.2) and numerical features (Section 5.3), after which we inspect variation across genres (Section 5.4).

5.1 Feature Importance in Random Forest Classifiers

In order to investigate the relative importance of various linguistic features in the environments of bridging anaphora and antecedents, we train a random forest tree ensemble on the GUMBridge data to identify bridging anaphora mentions and bridging antecedent mentions, using several classification tasks. For the first, bridging anaphor detection (distinguishing whether a mention of an entity is a bridging anaphor), we use all of the bridging anaphora in GUMBridge and sample an equal number of randomly selected non-bridging mentions to create balanced classes. For bridging antecedent detection, we also sample to create balanced classes and train an equivalent classifier. In our classifiers, we use numerical features, such as document position and mention length, as well as categorical features, such as entity type and head part of speech, to explore the linguistic environments of bridging anaphor and antecedent mentions. Details on the full list of features used in these models are included in Appendix B. We use optuna[3] to conduct a parameter search over 50 trials for each of our classifiers, selecting the classifiers with the best performance on the development set for analysis.

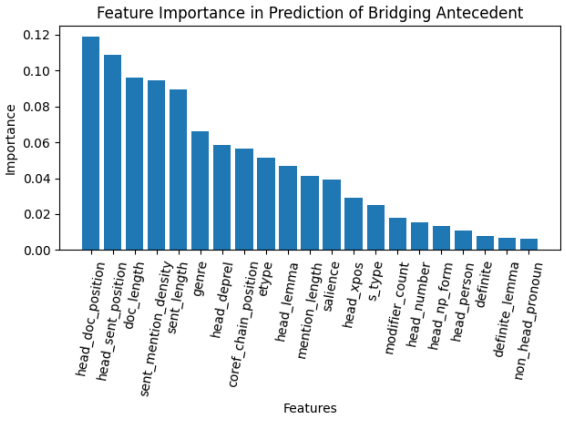

Figure 5: Feature importance for random forest classifier detecting bridging antecedent mentions.

The performance of the anaphor mention identification classifier and the antecedent mention identification classifier on class balanced samples of the held out GUMBridge test set are shown in Table 3. We see that the classifier performance on detecting bridging anaphora is higher than the performance on detecting bridging antecedents (F1 65.6 vs F1 54.8), indicating that the linguistic environments of bridging anaphora are more distinctive than their antecedents, which is not surprising, as there are more linguistic constraints on the definition of a bridging anaphor.

The feature importances of the anaphor detection classifier are shown in Figure 4 and the feature importances of the antecedent detection classifier are shown in Figure 5. Comparing the two figures, we see that the ordering of features is relatively similar, with positional numeric features at the top, followed by various categorical features including entity type, part of speech, and dependency relation. Notable differences include mention length and definiteness being ranked higher for anaphor

[3]https://github.com/optuna/optuna

Mention Type	Entity Type				Part of Speech				Dependency Relation			
	Pos		Neg		Pos		Neg		Pos		Neg	
Anaphor	time	**16.40**	person	**-19.08**	NNS	**27.98**	PRP	**-28.91**	obl	**18.09**	nmod:poss	**-13.81**
Antecedent	place	15.19	person	-10.51	NNPS	4.83	PRP	-6.05	obl:unmarked	8.51	conj	-8.76
Nonbridging	person	7.79	time	-5.95	PRP	9.49	NNS	-8.38	compound	4.15	obl	-6.40

Table 4: Strongest positive and negative χ^2 residual labels for entity type, part of speech, and syntactic dependency relations for bridging anaphor, bridging antecedent, and non-bridging mentions.

detection, while mention salience[4] ranks higher for antecedent detection. In the following section, we investigate how these features present for bridging anaphora, antecedents, and non-bridging mentions/entities.

5.2 Categorical Features: Entity Types, Part of Speech, and Dependency Relations

In Table 4, we show the strongest positive and negative χ^2 residual labels for entity type, part of speech, and syntactic dependency relations for bridging anaphor, bridging antecedent, and non-bridging mentions.[5] The entity types are from the GUM tag set (*person, place, organization, object, event, time, substance, animal, plant, abstract*), the part of speech labels are of the Penn Treebank tag set (Santorini, 1990), and the syntactic annotations are in the Universal Dependencies formalism.

First, if we look at the entity type residuals on the left of Table 4, we see that *person* entities are favored by non-bridging mentions and disfavored by both bridging anaphora and antecedents. While there are certainly instances where *person* entities can be a bridging anaphor or antecedent (e.g., relations nouns such as a parent → **the child**), *person* entities are very common and can frequently be understood independently upon first introduction by use of a name (proper noun) or a deictic expression (e.g., "you"). Looking at the positive entity type residuals, we see that *time* entities are associated with bridging anaphora and *place* entities are associated with bridging antecedents. Bridging anaphora that are time entities largely correspond to instances of COMPARISON-TIME and SET-SPAN-INTERVAL (e.g., today → **this morning**), and place entities as antecedents are common anchors for relative/associated entities (e.g., York → **the south** (of York)).

Next, if we look at the part of speech residuals in the center of Table 4, we see that pronouns (PRP) are favored by non-bridging mentions and disfavored by both bridging anaphora and antecedents. As pronouns are very frequently *person* entities, this mirrors our previous finding that *person* entities are less likely to be part of a bridging pair. Looking at the positive part of speech residuals, we see that plural nouns (NNS) are associated with bridging anaphora and plural proper nouns (NNPS) are associated with bridging antecedents. This suggests, unsurprisingly, that bridging entities are typically common nouns, and that proper nouns, which may be salient to the discourse (e.g., place names like "York"), can serve as anchors for bridging relations.

Finally, if we look at the dependency relation residuals on the right of Table 4, we see that oblique arguments/modifiers (*obl* and *obl:unmarked*), which include temporal modifiers characteristic of the COMPARISON-TIME subtype, are associated with bridging entities. We see that possessors (*nmod:poss*), which include possessive pronouns that are typically subsequent mentions (e.g., "his"), are not associated with bridging anaphora. We also find that compound and conjunction (*conj*) are positively associated with non-bridging mentions and negatively associated with bridging antecedents, respectively. This makes sense, as both of these relations suggest a more direct relation to another entity than an associative bridging relation. For instance, in "cats *and* dogs" the entities are grouped together via the explicit conjunction *and*, rather than relying on an implicit, anaphoric relation between the entities.

5.3 Numerical Features: Mention Length, Cluster Size, Salience, and Pair Distance

In this section, we explore several numerical features for bridging anaphor entities, bridging antecedent entities, and nonbridging entities. For mention length, we examine the length (in tokens) of the first mention of each entity. Cluster size refers to the number of mentions an entity has in a document, and salience refers to an entity's graded

[4]Salience is annotated in the corpus based on the number of times an entity is mentioned in multiple summaries of each document, see Lin and Zeldes (2025); Zeldes and Lin (2026).

[5]The full residual tables are shown in Appendix C.

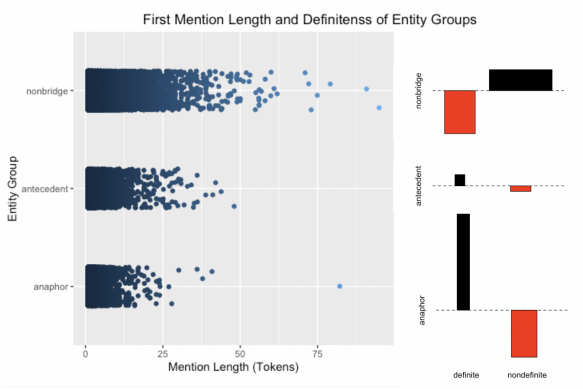

Figure 6: First mention length (left) and χ^2 residuals for definiteness (right) of anaphor, antecedent, and non-bridge entities.

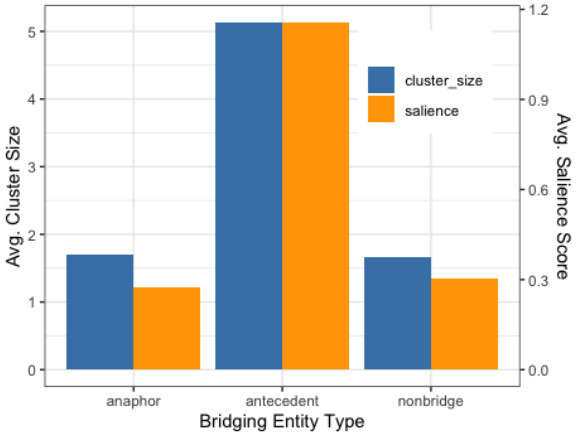

Figure 7: Average coref cluster size and salience score of anaphor, antecedent, and nonbridge entities.

entity score (0-5) (Lin and Zeldes, 2025), which reflects how prominent an entity is in a document. For pair distance, we examine the token distances between the anaphor and antecedent mentions of a bridging pair.

In Figure 6 we show the first mention length (left) and χ^2 residuals for definiteness (right) of anaphor, antecedent, and non-bridging entities. We can see that non-bridging entities tend to be longer, while bridging anaphora are typically shorter. This mirrors the tendency for subsequent mentions in coreference clusters (anaphora of identity relations) to be shorter than the initial mention (e.g., "a man with a top hat" will likely be referred to as "he" in a subsequent mention; adding more/longer descriptors in subsequent mentions is less likely). We note, however, that it is possible for bridging to form chains, despite the non-transitive nature of the phenomenon (Miéville, 1999): an anaphor may become an antecedent, as in the house → **the door** → **the handle** (but not #the house → **the handle**).

Looking at the right side of the figure, we see that there is also a strong tendency for bridging anaphora to be definite when compared to non-bridging entities. In English, a definite first mention often means that a participant can infer a specific referent despite the entity being newly introduced, creating a strong signal for a potential bridging anaphor. Definite bridging anaphora have been the main focus of earlier works on bridging (e.g, Poesio et al., 1997) because definite first mentions are relatively straightforward to identify, though we stress that indefinite bridging is possible, as in (4).

(4) The Sea Org requires **members** to sign a billion-year contract.

It is notable that even though GUMBridge is not restricted to definite bridging anaphora, there is still an observable tendency for bridging anaphora to be definite.

In Figure 7 we show the average cluster size (blue) and average salience score (orange) of anaphor, antecedent, and non-bridge entities. We can see quite clearly that bridging antecedents are part of larger entity clusters and are more salient than bridging anaphora and other nonbridging entities. The combination of these two findings indicates that a good candidate for a bridging antecedent will be an entity that occurs prominently in a discourse, which is therefore easier for participants to refer back to when resolving a bridging relation. This relates to Von Heusinger and Schumacher (2019)'s notion that prominence can act as an attractor of various linguistic operations due to higher levels of cognitive activation. Because prominent entities have a higher rate of recurrence, and thus give rise to referential continuity, we believe they are easier for discourse participants to refer back to and have the potential to act as anchors for the construction of bridging relations, a finding we believe has not been demonstrated to date.[6]

In Figure 8 we show the distribution of anaphor-antecedent distances for bridging pairs in GUMBridge grouped by whether the antecedent of the pair is a salient entity (graded salience score > 0). We can see that the distance between bridging anaphora and their antecedents is typically rel-

[6]For a possibly related finding, see Lin and Zeldes (2026), who show that salient entities make their surrounding context more predictable than non-salient ones.

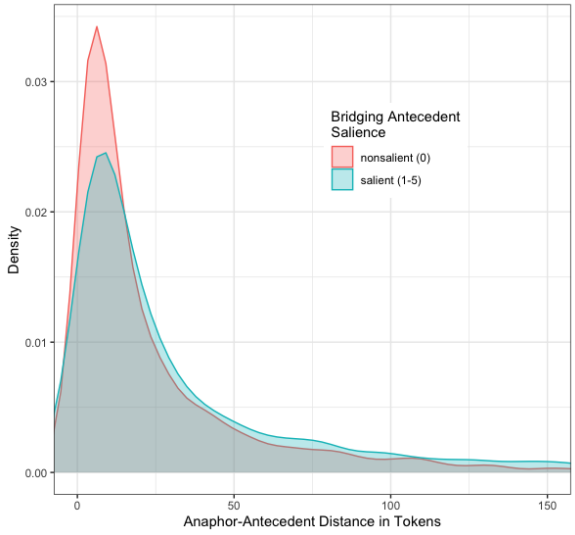

Figure 8: Distribution of anaphor-antecedent distances for bridging pairs for pairs with salient (blue) and non-salient (red) antecedents.

atively small (> 90% of instances are < 150 tokens apart[7]), indicating that bridging is predominantly a short range phenomenon. However, there is still a long tail of bridging instances beyond a distance of 150 tokens, which shows that long distance bridging does occur. We separate out instances with salient and non-salient antecedents in order to determine whether the antecedents of long distance bridging instances have a tendency to be more salient. We can see in Figure 8 that a larger part of the density curve for salient entities covers higher token distances, and a Kolmogorov-Smirnov test of the two distributions confirms that salient antecedents correspond with larger distances between the anaphor and antecedent of the bridging pair (p-value = 3.722e-12). Building on our previous finding that bridging antecedents are generally more likely to be salient than other entities, this suggests that more prominent entities are also better suited to act as anchors for bridging relations over long distances.

5.4 Cross-Genre Analysis of Numerical Features

In this section, we explore how the first mention length, cluster size, and salience of bridging entities (anaphora and antecedents) vary across genres. In order to examine how these environmental features vary beyond entity norms for each genre, we report mean z-scores of bridging anaphora and antecedent properties for each genre in GUMBridge.

[7]The long tail of the plot is cut off for ease of viewing.

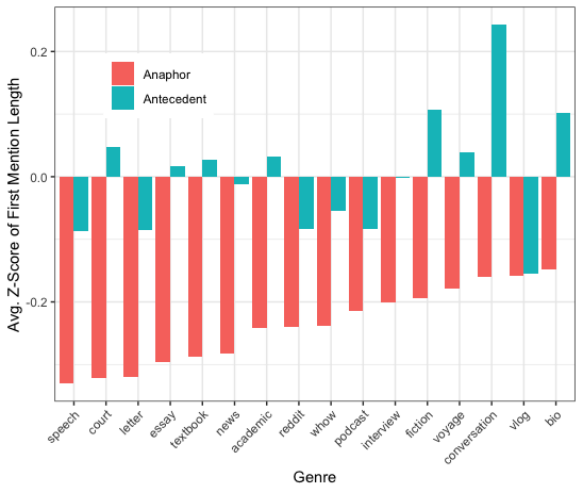

Figure 9: Average z-score of first mention length of bridging anaphora and antecedent by genre.

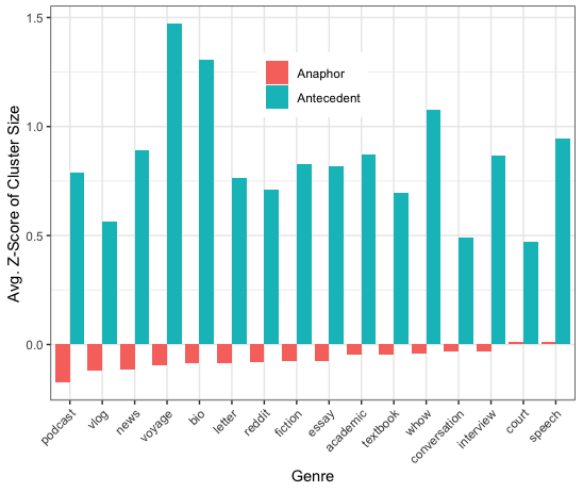

Figure 10: Average z-score of cluster size of bridging anaphora and antecedent by genre.

In Figure 9, we show the average z-scores of the first mention length (in tokens) of bridging anaphor and antecedent entities by genre. We see that the first mentions of bridging anaphora are uniformly shorter than the average entity across all genres, while the first mention length of bridging antecedents does not show a particular pattern. This indicates that regardless of genre, bridging anaphora display a tendency to be shorter than other entity mentions, which is similar to other varieties of anaphora outside of bridging.

In Figure 10, we show the average z-scores of the entity cluster size of bridging anaphor and antecedent entities by genre. Across all genres, we see that there is a clear tendency for the cluster size of bridging antecedents to be larger than the aver-

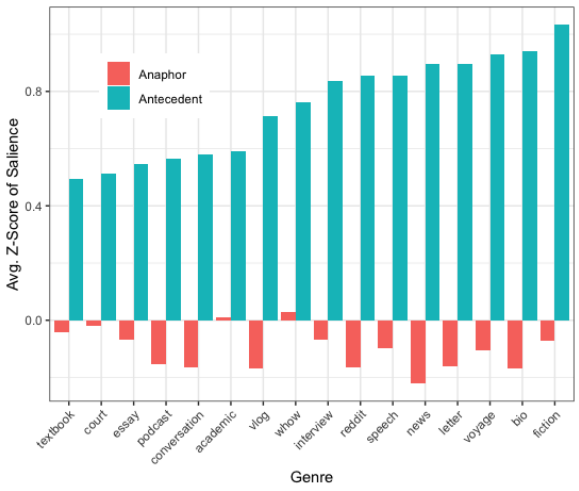

Figure 11: Average z-score salience score of bridging anaphora and antecedent by genre.

age entity. We note that the two genres with the highest z-scores, *voyage* and *bio*, are genres where a single entity (a *place* or a *person* respectively) is the main focus of the document, and is thus likely to be a prominent entity in the discourse. For most genres, we also see that bridging anaphora tend to have smaller cluster sizes than the average entity, though the effect of the trend is less than the one observed for bridging antecedents. For the two genres which have positive z-scores for anaphor cluster size (*court* and *speech*), the deviation from the average is very slight (both ~0.01). Overall, this lack of repeated mentions indicates that bridging anaphora are not generally discussed at length after their initial introduction.

In Figure 11, we show the average z-scores of the entity salience of bridging anaphor and antecedent entities by genre. We see that across all genres there is a tendency for bridging antecedents to be more salient than the average entity. We note that of the 3 genres with the highest z-scores, *voyage*, *bio*, and *fiction*, two of them (*bio* and *voyage*) were also the top genre for antecedent cluster size. For anaphora, we see that most genres show a slight tendency for bridging anaphora to be less salient than the average entity. This coincides with the previous finding that bridging anaphora are typically not the focus of the discussion following their introduction (as indicated by not having larger than average cluster sizes).

6 Conclusion

In this paper, we conducted a descriptive corpus analysis on the varieties of bridging anaphora in the GUMBridge corpus. This is the first multi-genre corpus analysis conducted for referential bridging in English, and we report a variety of novel insights regarding the behavior of bridging anaphora: (1) Looking at the distribution of bridging instances in the corpus, we observed that spoken genres have less bridging instances than written ones, and that ENTITY-ASSOCIATIVE and COMPARISON relations are the most attested forms of bridging across the full data set. (2) In our analysis of the categorical features in bridging environments, we observed a tendency for bridging entities to ground the setting of a discourse, i.e. *time* and *place*, rather than *person* entities. (3) In our analysis of numerical features, we observed that bridging anaphora have a tendency to have shorter first mentions and their first mentions also show a greater tendency to be definite when compared with nonbridging entities. (4) We also observed that bridging antecedents show a tendency to be from larger entity clusters and be more salient than other entities, and also that long distance instances of bridging are even more likely to have a salient entity as the antecedent of the bridging pair. (5) Finally, in the cross-genre comparison, we saw confirmation that observed trends in the first mention length of bridging anaphora, and cluster size and salience of bridging antecedents are consistent across genres.

Limitations

The corpus analysis conducted in this paper is limited to the 16 genres of English represented in the main partitions (train, dev, test) of the GUMBridge corpus. It is also limited to analyzing the phenomenon of bridging via the information status based definition of referential bridging and the subtype schema that was used during the construction of the GUMBridge corpus. As such, the findings may not generalize to resources which use significantly different definitions of bridging in their construction.

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.

Nicholas Asher. 1993. *Reference to abstract objects in discourse*. Kluwer, Dordrecht.

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3222–3228, Reykjavik, Iceland.

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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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.

John A. Hawkins. 1978. Definiteness and indefiniteness: A study in reference and grammaticality prediction. *Journal of Linguistics*, 27:405–442.

Lauren Levine and Amir Zeldes. 2026. GUMBridge: A corpus for varieties of bridging anaphora. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6823–6837, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Jessica Lin and Amir Zeldes. 2025. GUM-SAGE: A novel dataset and approach for graded entity salience prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 438–455, Vienna, Austria. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2026. Expect the unexpected? Testing the surprisal of salient entities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, San Diego, CA. 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.

Denis Miéville. 1999. Associative anaphora: An attempt at a formalization. *Journal of Pragmatics*, 31:327–337.

Anna Nedoluzhko, Jiří Mírovský, Radek Ocelák, and Jiří Pergler. 2009. Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, pages 1–16, Goa, India.

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).

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).

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

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

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).

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.

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.

Klaus Von Heusinger and Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127.

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

Amir Zeldes. 2022. Opinion piece: Can we fix the scope for coreference? *Dialogue & Discourse*, 13:41–62.

Amir Zeldes and Jessica Lin. 2026. What makes an entity salient in discourse? *Corpus Linguistics and Linguistic Theory*, pages 1–42.

A GUMBridge Bridging Subtypes

This appendix briefly details the bridging subtype varieties annotated in the GUMBridge corpus, which are reflected in its guidelines. For the full guidelines, we refer readers to Levine and Zeldes 2026.

COMPARISON-RELATIVE The anaphor is preceded by a comparative marker which implies a comparison to the antecedent (e.g., several women → **other women**).

COMPARISON-SENSE The type of the anaphor is omitted but inferable via comparison to the antecedent (e.g., a Chinese restaurant → **the Italian one**; "one" is of type "restaurant").

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 (e.g., Wednesday → **yesterday**).

ENTITY-MERONOMY The anaphor has a part-whole relation with the antecedent, including physical subparts, substance-portion, and regions/subsections (e.g., a house → **the door**).

ENTITY-PROPERTY The anaphor is a physical or intangible property of the antecedent, such as smell, length, or style (e.g., a bouquet of roses → **the scent**).

ENTITY-RESULTATIVE The anaphor is logically inferable from the antecedent. This is often the result of a transformative/product producing process, like cooking/baking (e.g., flour → **the bread**).

ENTITY-ASSOCIATIVE The anaphor is an attribute or closely associated entity of the antecedent (e.g., a library → **the books**).

SET-MEMBER The anaphor is an element of the antecedent set. This includes group-member and class-instance relations (e.g., several books → **the mystery novel**).

SET-SUBSET The anaphor is a subset of the antecedent set (e.g., a group of students → **the boys**).

SET-SPAN-INTERVAL The anaphor is a sub-span of the spatial or temporal antecedent interval (e.g, Sunday → **the morning**).

OTHER The OTHER category is for instances which fit the information status based definition of a bridging pair but do not fall into any of the bridging subtype categories outlined above.

B Description of Features used in Random Forest Classifiers

Below are the features used in our random forest classifiers to distinguish bridging anaphor/antecedent mentions from other mentions. These features are derived from the various linguistic annotations in the GUM corpus (Zeldes, 2017):

salience: Graded salience score (0-5) of the entity that the mention is an instance of.

head_doc_position: Document token position of the head token of the mention.

mention_length: Number of tokens in the mention.

definite: Binary indication of whether the mention is definite.

etype: GUM style entity type of mention (e.g., *abstract*, *animal*, etc.).

genre: Genre of the document that the mention is from (e.g., *academic*, *bio*, etc.).

head_lemma: Lemma of the head token of the mention.

head_deprel: UD syntactic dependency relation of the head token of the mention.

head_xpos: Part of speech (Penn Treebank tag set) of the head token of the mention.

head_number: Number (e.g., *Sing*) of the head token of the mention.

head_np_form: Noun phrase (NP) form of the head token of the mention, selecting from *pronoun* (PRP, PRP$), *proper* (NNP, NNPS), *noun* (NN, NNS), and *other* (all other parts of speech).

doc_length: Length of document that the mention is in (in tokens).

modifier_count: Count of modifier dependency relations (*mod*, *compound*, *relcl*, *nmod*, *nmod:poss*) within the mention that have the head token of the mention as their head.

Entity Type	Anaphor	Antecedent	Nonbridge
abstract	5.51	-5.14	-0.34
animal	-2.02	1.79	0.15
event	-0.04	2.40	-0.55
object	10.11	3.66	-3.68
organization	0.27	3.42	-0.88
person	-19.08	-10.51	7.79
place	4.06	15.19	-4.68
plant	1.57	1.26	-0.73
substance	-2.03	-0.28	0.63
time	16.40	5.85	-5.95

Table 5: Full χ^2 residuals for entity type and mention type (bridging anaphor, bridging antecedent, or nonbridging mention). (X-squared = 1364.1, df = 18, p-value < 2.2e-16)

sent_length: Length of the sentence the mention is in (in tokens).

head_sent_position: Relative position of the head token of the mention in the sentence (0-1).

sent_mention_density: Count of mentions in sentence / length of sentence (in tokens).

non_head_pronoun: Binary indication of whether there is a pronoun besides the head token in the mention.

coref_chain_position: Relative position of the mention in its coref chain (0-1) (for antecedent classifier only).

s_type: Type of sentence the mention is in (e.g., *decl* for declarative).

head_person: Person (e.g., *first*) of the head token of the mention.

definite_lemma: Lemma of the token in the mention with dependency relation *det* (*null/unknown* if there is no such token).

C Full χ^2 Residuals for Categorical Features

In this appendix, we give the full χ^2 residuals for entity type, part of speech (of the mention head), and syntactic dependency relation (of the mention head) across mention types (bridging anaphor, bridging antecedent, and nonbridging mention). Table 5 gives the residuals for entity types, Table 6 gives the residuals for parts of speech, and Table 7 gives the residuals for syntactic dependency relations.

POS	Anaphor	Antecedent	Nonbridge
$	-1.67	-1.59	0.84
CD	-2.14	1.18	0.32
DT	-1.83	-1.48	0.86
FW	1.27	-1.51	-0.001
JJ	11.44	-0.75	-3.02
JJS	2.51	-1.67	-0.31
NN	20.82	1.73	-6.22
NNP	-18.93	2.05	4.81
NNPS	-4.79	4.83	0.21
NNS	27.98	2.41	-8.38
PRP	-28.91	-6.05	9.49
PRP$	-14.82	-1.20	4.42
RB	5.90	2.86	-2.32
VB	-6.01	-1.68	2.07
VBD	-3.63	-0.40	1.11
VBG	-3.58	-1.07	1.25
VBN	-2.05	-0.81	0.76
VBP	-2.89	-0.45	0.91
VBZ	-2.86	0.11	0.77
WP	-0.93	-1.16	0.53

Table 6: Full χ^2 residuals for part of speech (POS; Penn Treebank tag set) of the head token of the mention and the mention type (bridging anaphor, bridging antecedent, or nonbridging mention). POS labels were limited to those with at least 50 occurrences. (X-squared = 3288.3, df = 38, p-value < 2.2e-16)

Deprel	Anaphor	Antecedent	Nonbridge
acl	-1.92	-0.89	0.74
acl:relcl	-2.12	-1.94	1.05
advcl	-1.39	0.51	0.27
advmod	10.50	2.20	-3.45
appos	-7.56	4.18	1.14
ccomp	-3.40	-0.77	1.13
compound	-11.15	-3.70	3.99
conj	5.53	-8.28	0.39
csubj	-4.64	-1.82	1.73
dep	-2.56	-1.21	1.00
dislocated	-1.03	-0.91	0.50
expl	-4.49	-3.76	2.14
iobj	-3.44	-0.55	1.09
list	-2.68	-1.35	1.07
nmod	6.52	1.91	-2.27
nmod:poss	-14.01	0.59	3.78
nmod:unmarked	-2.20	1.39	0.29
nsubj	-11.50	-2.17	3.72
nsubj:outer	5.13	-1.43	-1.10
nsubj:pass	4.11	0.95	-1.37
obj	6.10	3.32	-2.48
obl	20.13	3.49	-6.45
obl:agent	-0.05	-0.68	0.17
obl:unmarked	9.85	5.62	-4.07
orphan	2.03	-1.07	-0.32
parataxis	-4.64	-2.80	1.95
reparandum	-1.18	-2.22	0.85
root	-2.90	1.69	0.41
vocative	-3.20	-1.84	1.33
xcomp	-3.45	0.86	0.77

Table 7: Full χ^2 residuals for part of syntactic dependency relations (deprel) of the head token of the mention and the mention type (bridging anaphor, bridging antecedent, or nonbridging mention). Deprel labels were limited to those with at least 50 occurrences. (X-squared = 1801.4, df = 58, p-value < 2.2e-16)