Challenges to Evaluating the Generalization of Coreference Resolution Models: A Measurement Modeling Perspective

Ian Porada[1], Alexandra Olteanu[2], Kaheer Suleman[*], Adam Trischler[*], and Jackie Chi Kit Cheung[1,3]

[1]Mila, McGill University
ian.porada@mail.mcgill.ca, jackie.cheung@mcgill.ca
[2]Microsoft Research Montréal
alexandra.olteanu@microsoft.com
[3]Canada CIFAR AI Chair

Abstract

It is increasingly common to evaluate the same coreference resolution (CR) model on multiple datasets. Do these multi-dataset evaluations allow us to draw meaningful conclusions about model generalization? Or, do they rather reflect the idiosyncrasies of a particular experimental setup (e.g., the specific datasets used)? To study this, we view evaluation through the lens of measurement modeling, a framework commonly used in the social sciences for analyzing the validity of measurements. By taking this perspective, we show how multi-dataset evaluations risk conflating different factors concerning what, precisely, is being measured. This in turn makes it difficult to draw more generalizable conclusions from these evaluations. For instance, we show that across seven datasets, measurements intended to reflect CR model generalization are often correlated with differences in both how coreference is defined and how it is operationalized; this limits our ability to draw conclusions regarding the ability of CR models to generalize across any singular dimension. We believe the measurement modeling framework provides the needed vocabulary for discussing challenges surrounding what is actually being measured by CR evaluations.

1 Introduction

Coreference is the broad phenomenon of multiple linguistic expressions referring to the same discourse entity (see examples in Table 1). Coreference resolution (CR) is then the task of identifying those expressions that are coreferring (Winograd, 1972; Kantor, 1977; Hirst, 1981, *i.a.*). This task has been studied extensively (*e.g.,* Sukthanker et al., 2020; Poesio et al., 2023) and is considered a core part of representing the semantics of natural language (Hobbs, 1978).

Dataset	Example
OntoNotes	Highway officials insist the ornamental railings on older bridges aren't strong enough to prevent vehicles from crashing through. But other people don't want to lose the bridges' beautiful, sometimes historic, features.
PreCo	"Melting permafrost can also destroy trees and forests," Weller said. "When holes in the ground form, trees fall into them and die."
Phrase Detectives	Within 1942-1944 bicycles were also added to regimental equipment pools. … Although seeing heavy use in World War I, bicycles were largely superseded by motorized transport in more modern armies.

Table 1: Examples of annotated coreference (highlighted) where at least one expression could be interpreted as generic. Generic expressions tend to be annotated differently across datasets in part due to varying definitions of coreference.

As a result, a common goal has been to build CR models that are able to generalize across settings (Urbizu et al., 2019; Xia and Van Durme, 2021; Žabokrtský et al., 2022). This goal has motivated the design of benchmarks "to measure progress towards truly general-purpose coreference resolution" (Toshniwal et al., 2021), and has led to CR models being increasingly evaluated across multiple datasets (*e.g.,* Yang et al., 2012; Poot and van Cranenburgh, 2020; Straka, 2023).

However, *what we can learn from multi-dataset evaluations about a CR model's ability to generalize will necessarily be limited by differences in how the datasets were constructed.*

Consider the following example: researcher **A** defines coreference to be a relationship exclusively between noun phrases and uses crowdworkers to annotate a dataset based on this definition. Independently, researcher **B** defines coreference to be a relationship between any noun phrase or verb phrase and annotates a different dataset based on this looser definition. A CR model is then trained using **A**'s dataset and evaluated on **B**'s. Without even seeing the evaluation results, one might already suspect that what we can learn from those results about how well a model trained on one dataset generalizes to the other will depend on the differ-

---

ences in how coreference was defined and annotated across the two datasets.

We aim to unpack how differences in the definition and annotation of coreference across datasets might limit the conclusions that can be drawn about the generalization ability of CR models. To do so, we examine common evaluation practices using the *measurement modeling* framework (Bollen, 2001; Jacobs and Wallach, 2021; Blodgett et al., 2021, *i.a.*). This framework helps us distinguish between what is being measured, the *construct*, from how it is being measured, the *operationalization* of the construct via a *measurement model*. This distinction is particularly important when working with contested constructs—i.e., constructs with disputed or many competing definitions—which latent constructs often are, as it helps differentiate between differences in the definitions of a construct versus differences in operationalizations of the same definition.

In fact, coreference is itself a contested construct (Grishman and Sundheim, 1996; Doddington et al., 2004; Zeldes, 2022); "[a] very basic problem arising in the case of coreference is deciding what type of information is being annotated, since the term 'coreference' is used to indicate different things" (Poesio et al., 1999). The measurement modeling framework thus helps us illustrate how common multi-dataset evaluations risk conflating different factors concerning what, precisely, is being measured; and, as a result, what we might be learning from these evaluations about models' ability to generalize across datasets.

In this work, we conduct a multi-dataset evaluation across seven English-language datasets of coreference annotations and analyze the results through the lens of measurement modeling. Specifically, we consider how coreference is defined and annotated within each dataset (§3), and how this affects measurements of model generalization (§4 & §5). For this, we identify particular types of coreference that vary between definitions—e.g., the coreference of generic and predicative expressions—and show, using disaggregated results, how models systematically fail in cases where the definition of coreference used to build the test set differs from that used for the training set.

For instance, consider the coreference annotated in PreCo and highlighted in Table 1. A state-of-the-art CR model trained on OntoNotes fails to predict the two instances of "trees" as coreferring. This may be attributable to the fact that most often two

generic nouns are not considered coreferring under the definition of coreference used in OntoNotes.

Our results suggest that to understand if and how CR models generalize some concept of coreference requires us to *first* resolve inconsistencies in how coreference is defined. More broadly, through disaggregated evaluations and by applying the measurement modeling framework, our work provides a blueprint for evaluating out-of-domain performance in a way that accounts for known inconsistencies in how coreference is defined and operationalized across datasets.

2 Background and Related Work

CR Model Error Analyses. Our work relates to efforts to identify and classify model errors. Error analyses of CR models, however, have focused on performance measurements based on a single dataset (Uryupina, 2008; Versley, 2008; Durrett and Klein, 2013). Analyses that consider collections of datasets find that many types of errors are consistent across datasets, but often do not examine cases where models are tested on datasets they were not trained on (Lu and Ng, 2020; Chai and Strube, 2023). Furthermore, these prior studies evaluating CR models' performance across datasets—as well as studies that analyze the models' downstream performance (Dasigi et al., 2019; Chai et al., 2022)—have not considered how aspects related the way in which coreference is defined might correlate with the observed errors.

CR Model Generalization. To improve generalization, several approaches aim to train a CR model using multiple training sets (*e.g.,* Zhao and Ng, 2014; Gandhi et al., 2022; Yuan et al., 2022). Such work has focused mostly on *active learning*. It has been observed that models trained on multiple datasets do not improve in accuracy on a test set for which large amounts of in-domain training data is available (Toshniwal et al., 2021). Related to our work, Moosavi et al. (2019) studied how coreference annotations differ in terms of the span boundaries annotated and proposed an algorithm for heuristically determining the minimal span boundary of a mention.

Models trained on OntoNotes are believed to be less accurate at CR for textual genres and for proper nouns not in the training set (Moosavi and Strube, 2017; Subramanian and Roth, 2019; Zhu et al., 2021). Prior work has improved on existing generalization measurements in these cases by

incorporating explicit linguistic features or syntactic rules (Zeldes and Zhang, 2016; Moosavi and Strube, 2018). Prior work has also examined the generalization of CR models to supposedly difficult cases of pronominal coreference (Rahman and Ng, 2012; Peng et al., 2015; Toshniwal et al., 2021). For instance, models trained on OntoNotes are believed to generalize poorly to examples of pronominal coreference scraped from the web (Webster et al., 2018; Emami et al., 2019).

Measurement Modeling. How to effectively measure theoretical concepts has been studied extensively in the quantitative social sciences (*e.g.,* Black, 1999; Adcock and Collier, 2001; Bhattacherjee, 2012). Because theoretical concepts are often unobservable and can therefore not be measured directly, they must be inferred based on observable variables or the inferred measurements of other theoretical concepts. A model that describes the relationship between a theoretical concept and the variables from which that concept is inferred is called a *measurement model* (Bollen, 2001; Jacobs and Wallach, 2021). Typically, the term *construct* is used to refer to the theoretical concept being measured and the term *indicator* is used to refer to any of the observable variables from which the measurement of the construct is inferred (Kline, 2011).

We draw on these lines of work to examine how decisions made when constructing CR evaluation datasets might limit the type of conclusions we can draw about models' ability to generalize from one dataset to another. Specifically, the measurement modeling framework can help us examine how factors related to how coreference is defined and annotated might affect the measurement of the constructs of *coreference*, *CR model performance*, or *CR model generalization*.

3 Coreference

Coreference is itself an unobservable construct; therefore, measurements of coreference must be inferred from indicators, e.g., annotations of coreference in a dataset.

In this section, we first describe the coreference datasets we consider in our analysis (§3.1). We then describe discrepancies in how coreference is defined (§3.2.1) and operationalized (§3.2.2) within these datasets.

Dataset	Train	Dev.	Test	Total Words (K)
OntoNotes	2,802	343	348	1,632
PreCo	36,120	500	500	12,493
Phrase Det.	695	45	45	1,321
OntoGUM	165	24	24	204
LitBank	80	10	10	211
ARRAU	444	33	75	348
MMC	955	134	133	324

Table 2: Number of documents for each dataset split.

3.1 Datasets

Our goal is to understand whether multi-dataset evaluations tell us anything about models' ability to generalize. To examine this, we focus our analysis on datasets that have been used in prior work to evaluate CR model generalization. Specifically, our selection of datasets is based on prior multi-dataset evaluations of CR models (Toshniwal et al., 2021; Xia and Van Durme, 2021; Zhu et al., 2021; Žabokrtský et al., 2023). In total, we consider seven English-language datasets containing annotations of identity coreference at a document level (see Table 2 for dataset figures). Appendix A.1 provides more details regarding the selection of these datasets, how they have been preprocessed or formatted, and the use of these datasets within the broader coreference literature.

OntoNotes. OntoNotes 5.0 (Weischedel et al., 2013) consists of news, conversations, web data, and biblical text in which coreference was annotated by experts. We use the English CoNLL-2012 Shared Task version of this dataset (Pradhan et al., 2012). While OntoNotes 5.0 is annotated for multiple phenomena, the CoNLL-2012 Shared Task contains only so-called "identical" coreference, also sometimes referred to as "anaphoric coreference."

PreCo. PreCo (Chen et al., 2018) consists of English comprehensive exams annotated for coreference by trained university students.

Phrase Detectives. Phrase Detectives 3.0 (Yu et al., 2023b) consists of Wikipedia, fiction, and technical text. The training set annotations were sourced by aggregating annotations of users playing the *Phrase Detectives* online game, where the users were tasked with annotating coreferences, or verifying others' annotations. The test set was annotated by experts. We use the "CoNLL" formatted version of the dataset.[1]

OntoGUM. OntoGUM (Zhu et al., 2021) is a reformatted version of the GUM corpus (Zeldes, 2017). The GUM corpus was originally annotated

---

[1] https://github.com/dali-ambiguity/Phrase-Detectives-Corpus-3.0

15382

in an iterative process by linguistics students. OntoGUM was created by transforming GUM, using deterministic rules and annotated syntactic parses, to follow the OntoNotes annotation guidelines. We use version 9.2.0 of OntoGUM.

LitBank. LitBank (Bamman et al., 2020) consists of coreference annotated in English literature by experts. Only noun phrases of ACE categories and pronouns have been annotated for coreference.[2] We use the "CoNLL" formatted dataset version.[3]

ARRAU. ARRAU 2.1 (Uryupina et al., 2020) is a dataset of written news and spoken conversations annotated for various anaphoric phenomenon by experts. We use all documents and the formatting procedure of Xia and Van Durme (2021) which keeps only the coarsest-grained annotations.

MMC Multilingual Coreference Resolution in Multiparty Dialogue (MMC) (Zheng et al., 2023) is a dataset of television transcripts. The training set was annotated by crowdworkers, and the test set by experts. We use the English portion of the "CoNLL" formatted version of this dataset.[4]

3.2 Differences Across Datasets

Here, we overview differences in how coreference is defined and operationalized across the three largest datasets—OntoNotes, PreCo, and Phrase Detectives—which we also use as training sets.

3.2.1 Differing Definitions

Because datasets differ in how coreference is conceptualized, measurements of how CR models generalize from one dataset to another might be confounded by or rather capture the differences in the corresponding definitions of coreference.

Notable differences in how these three datasets conceptualize coreference primarily stem from whether the following phenomena are considered to be coreference relations: **1)** multiple **generic** expressions that could possibly be interpreted as referring to the same discourse entity (i.e., generic only), **2)** expressions that could be interpreted as referring to the same event where at least one is a **verb phrase** (VPs), **3)** two expressions in **apposition**, and **4)** two expressions in a **copular structure**. We overview whether datasets include these in the definition of coreference in Table 3.

These differences in how coreference is defined were determined based on the original documentation of the respective datasets. We considered both the original publication as well as the annotation guidelines where available. More details on the definitions of coreference for each dataset are provided in Appendix A.2.

3.2.2 Differing Annotations

In addition to differences in how coreference is conceptualized, we also consider aspects related to how the construct is operationalized.

While there are many dimensions to how coreference is operationalized within each dataset, we focus on differences in datasets' annotation guidelines, particularly concerning how coreferring expressions are specified to be annotated. We highlight two key ways in which annotation differs between the three training datasets which can be empirically studied: **1)** the annotation of **nested mentions** and **2)** the annotation of **compound modifiers**. We provide more details of the differences between datasets in Table 3.

The definition of coreference influences the operationalization, and therefore differences in annotations and definitions are closely coupled. For example, the inclusion of apposition in the definition of coreference will affect how nested mentions are then annotated.

In fact, in our analysis we often consider factors related to the operationalization and conceptualization of coreference jointly as these are often not explicitly delineated in the datasets' documentations. More generally, there is little or no clear distinction in the literature between decisions related to conceptualizing coreference—that is, defining the theoretical construct being measured—versus decisions related to the measurement of coreference—how a given definition is being operationalized.

For instance, certain decisions are explicitly made to increase accuracy in terms of annotator agreement; in the case of annotating verb phrases in OntoNotes "[o]nly the single-word head of the verb phrase is included in the span, even in cases where the entire verb phrase is the logical coreferent" (BBN Technologies, 2007). However, dimensions such as the handling of compound modifiers could be argued to be either a way to clarify how annotators should approach their task and thus improve annotator accuracy, or part of the definition of coreference itself. Clarifying this distinction is a community-wide research endeavor.

---

[2] The ACE categories are: people, facilities, geo-political entities, locations, vehicles, and organizations.

[3] https://github.com/dbamman/litbank

[4] https://github.com/boyuanzheng010/mmc

Type	OntoNotes	PreCo	Phrase Detectives
Generic Mentions [dogs] can bark	Generics mentions are only annotated when they corefer with a pronoun or determinate noun phrase, or when they occur in a news headline.	All generic noun phrases and modifiers can be annotated as coreferring.	All generic noun phrases can be annotated as coreferring.
Verb Phrases it will [grow]	The head of a verb phrase can be annotated as coreferring with a determinant noun phrase.	Not annotated.	Not annotated.
Appositives [[Abe] , [the chef]]	Annotated in the dataset, but not considered coreference.	Annotated as three mentions: both noun phrases and the larger span.	Annotated in the dataset, but not considered coreference.
Copular Predicates [he] is [the teacher]	Not annotated.	In a copular structure, the referent and attribute are annotated as coreferring.	Annotated in the dataset, but not considered coreference.
Nesting [he [himself]]	When two nested mentions share a head, only the dominant mention is annotated. Proper nouns cannot contain nested mentions.	Appositives and mentions with shared heads are annotated as nested mentions.	The right-most mention in an appositive is considered referring to a distinct entity and can therefore be annotated as a nested mention. No restrictions on nesting of proper nouns.
Compound Modifiers [Taiwan] authorities	Compound modifiers are annotated if non-adjective proper nouns that are not a nationality acronym.	All compound modifiers can be annotated as coreferring.	No explicit restrictions on the annotation of compound modifiers.

Table 3: Noted differences in how coreference is defined and operationalized in the training datasets.

4 Methodology

This section first provides a high-level overview of the experimental setup (§4.1), and then describes the details of how the model evaluation is performed (§4.1.2), as well as the models we consider (§4.2). Our goal is to understand how measurements of generalization might be affected by differing definitions and operationalizations of coreference across datasets. To this end, we empirically examine performance variations for types of coreference that differ in how they are conceptualized and/or operationalized across datasets.

4.1 Disaggregated Evaluation

The purpose of disaggregated evaluations is to understand whether or not failures in generalization, as measured by a model's accuracy on multiple test sets, are correlated with differences in how CR is defined and operationalized in these test sets. For instance, if measurements of generalization seem indeed sensitive to differences in how coreference is defined across datasets, this might indicate that measurements of generalization capture or are confounded by differences in how coreference is defined rather than solely capturing the capacity of a CR model to generalize some consistent conceptualization of coreference.

We evaluate models on the test set of the dataset they were trained on, which we refer to as *in-domain*, as well as the test sets of all other datasets, which we refer to as *out-of-domain*. We then highlight types of coreference that appear significantly more difficult for a model to infer out-of-domain (different definitions or operationalizations) as compared to in-domain (same definition and opera-

tionalization). These cases are typically interpreted as indicating limited generalization, and we we want to examine whether such limited generalization is correlated with types of coreference that differ in the ways in which they are conceptualized and operationalized.

More formally, let θ be the parameterization of some model, X the in-domain dataset, $\hat{X}$ some out-of-domain dataset, and $f_\theta(X)$ some measurement of accuracy on dataset X. The *aggregate generalization gap* (AGG) is defined to be:

$$\text{AGG} = |f_\theta(X) - f_\theta(\hat{X})| \qquad (1)$$

Further, let X_t be the subset of dataset X that is of type t for which we can calculate model accuracy as $f_\theta(X_t)$. We will highlight cases where the aggregate generalization gap is significantly different from the *type generalization gap* (TGG) for a given type of coreference t, defined to be:

$$\text{TGG} = |f_\theta(X_t) - f_\theta(\hat{X}_t)| \qquad (2)$$

4.1.1 Formalizing Coreference Types

Here, we clarify the precise differences in how coreference is defined and operationalized that we consider in our study. We formalize these differences as particular types of coreference that vary in how they are handled between datasets. All of the differences in coreference annotation that we have noted are with respect to the annotation of particular types of mentions that occur in a coreference relation (Table 3), therefore our definition for each *type of coreference* is based on first formally defining particular *types of mentions*. For now, it is not important whether a given mention is annotated in the dataset or predicted by a model, as

we will clarify which is the case when performing evaluations of both precision and recall (§4.1.2).

Metadata. In practice, we identify some mention types using universal dependency relations (Nivre et al., 2020) or part-of-speech (POS) tags. This metadata is sourced as follows: for OntoNotes, we convert the expert annotated constituency parses into a dependency parse using the Stanford CoreNLP library (Manning et al., 2014). For On-toGUM, we use the existing expert annotated dependency parses. For PreCo, Phrase Detectives, and all other datasets, we parse the dataset using the Stanza parser (Qi et al., 2020).

Studied Coreference Types.

1. **Nested**: *A mention that corefers with an over-lapping span (either larger or smaller).*

> President Chen said, [he [himself]] has not re-turned to his hometown ...

2. **ON Generic**: *A coreferring mention that is a generic noun phrase.* We use the OntoNotes defini-tion of generic, which we refer to as an ON generic: a noun phrase is considered generic in OntoNotes if it has an indefinite article or is plural with no article. We check if there exists *a* or *an* as the child of a det relation, or there is no det relation and the POS tag is NNS (plural noun).

> ...you know yeah they had [a farm] when they were first married ...I don't know how many years they had [a dairy farm] ...

3. **Compound**: *A coreferring mention that is a compound modifier.* Defined to be a mention that is the dependent of a compound:nn relation.

> ... we miss our [Taiwan] compatriots even more, and ... gave a speech, expressing hopes that [Tai-wan] authorities would ...

4. **Copular**: *A mention that is in a copular con-struction with another mention to which it corefers.* Two mentions are said to be in a copular construc-tion if they are in an nsubj relation with each other and the rightmost mention is the head of a cop dependency relation.

> Yet I realize that in my father's eyes, [I] will al-ways be [his little girl].

4.1.2 Evaluation Metrics

In this section, we outline the specific metrics used to calculate the accuracy measurement $f_\theta(\cdot)$. To evaluate a model, we would like to know how well said model's predicted coreference clusters agree with the datasets' annotated clusters.

Notation. Let document $D = (w_1, w_2, \ldots, w_n)$ be a sequence of n words. Let $S = \{s_{i:j} \mid 1 \le i \le j \le n\}$ denote the set of all spans of words in D, where a span is a continuous subsequence of words $s_{i:j} = (w_i, \ldots, w_j)$. The set of annotated mentions $M \subseteq S$ denotes all spans in S that are annotated as coreferring with at least one other span. The set of entities $E = \{E_1, E_2, \ldots, E_k\}$ is a k-partition of M, i.e., a family of k non-empty, pairwise disjoint subsets of S. Each entity $E_i \in E$ represents a set of spans that refer to the same discourse entity in the document. A CR model takes as input D and outputs $E' = \{E'_1, \ldots, E'_\ell\}$, a prediction of the true E, which is an ℓ-partition of the set of model predicted mentions M'.

Metrics. There is not a single standard approach for calculating disaggregated CR evaluation scores. We choose to the adopt the B^3-based method of Bamman et al. (2020) due to its clear intuition, although in practice we find that our conclusions are not sensitive to the exact choice of metric.

In the aggregate case, we report agreement using B^3 recall, precision, and F1 (Bagga and Baldwin, 1998). The probabilistic intuition behind B^3 recall is that it calculates, in expectation over all anno-tated mentions, the probability that an annotated coreferring mention is correctly recalled by the model. A similar intuition applies to precision.

More formally, B^3 calculates cluster agreement by averaging over all mentions in the cluster. Given mention s, let $\mathcal{E} = \bigcup\{E_i \mid s \in E_i\}$ be the set of all mentions annotated as coreferring with s, and $\mathcal{E}' = \bigcup\{E'_i \mid s \in E'_i\}$ the set of all mentions predicted as coreferring with s. B^3 calculates recall (R) and precision (P) as

$$R = \frac{1}{|M|} \sum_{s \in M} \frac{|\mathcal{E} \cap \mathcal{E}'|}{|\mathcal{E}|} \qquad P = \frac{1}{|M'|} \sum_{s \in M'} \frac{|\mathcal{E}' \cap \mathcal{E}|}{|\mathcal{E}'|} \quad (3)$$

from which F1 is calculated as the harmonic mean.

To then evaluate performance on a given type of coreference, we use a modified B^3 score that calcu-lates the expected agreement only for mentions of the given type. That is, we only sum over mentions s that are of the type of mention being considered. In other words, M is replaced with $\{m \mid m \in M \land \text{is\_type}(m)\}$ when calculating recall, and M' is replaced with $\{m \mid m \in M' \land \text{is\_type}(m)\}$ when calculating precision. Where here, $\text{is\_type}(m)$ in-dicates that mention m is of the type being con-sidered. We also report CoNLL score (Denis and

Test \ Train	ON						PC			PD			
	LinkAppend			LingMess$_{ON}$			LingMess$_{PC}$			LingMess$_{PD}$			
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	# Ments.
OntoNotes (CoNLL)	83.7	82.9	83.3	79.2	82.6	80.9	64.1	48.8	55.4	33.5	61.4	43.3	
OntoNotes (B^3)	83.2	82.0	82.6	77.6	81.6	79.5	62.5	45.8	52.9	28.0	57.5	37.7	19,764
Nested	43.9	70.6	54.1	48.5	65.2	55.6	20.7	2.0	3.6	0.1	0.2	0.2	247
ON Generic	56.1	64.7	60.1	47.0	63.8	54.1	52.9	8.1	14.0	34.0	13.8	19.7	1,182
Compound	81.3	76.1	78.6	81.2	80.9	81.0	73.5	14.3	23.9	0.0	0.0	0.0	412
Copular	83.3	45.5	58.8	66.7	40.0	50.0	40.5	0.4	0.8	0.0	0.0	0.0	12
OntoGUM (CoNLL)	64.6	78.3	70.7	59.3	76.3	66.7	57.3	49.3	52.9	38.5	67.2	48.9	
OntoGUM (B^3)	65.0	75.6	69.9	59.3	73.6	65.7	58.3	46.0	51.4	33.0	61.8	43.0	2,707
Nested	25.0	2.1	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
ON Generic	34.5	50.0	40.8	18.4	23.4	20.4	64.4	3.9	7.3	21.0	3.5	6.0	169
Compound	50.6	69.2	58.1	35.4	74.3	47.5	48.7	13.5	21.2	0.0	0.0	0.0	64

Table 4: Aggregated and disaggregated metrics intended to measure model performance. Scores are correlated with differences in how coreference is defined and operationalized across datasets, indicating that purported measurements of out-of-domain generalization also encompass these differences between datasets. Each column corresponds to a model trained on the respective training set: OntoNotes (ON), PreCo (PC), or Phrase Detectives (PD). For LinkAppend (Bohnet et al., 2023), we use the publicly released weights. For LingMess (Otmazgin et al., 2022), we train the model on each of the three training sets. Each row corresponds to the specified type of coreference. F1 scores are highlighted for types of coreference where the F1 score dropped significantly out-of-domain as compared to the overall difference in aggregate scores. "# Ments." refers to the number of annotated mentions.

Baldridge, 2009; Pradhan et al., 2012) in the aggregate case to facilitate comparison with prior work.

4.2 Models

We focus our evaluation on four models: the publicly available LinkAppend model (Bohnet et al., 2023) and three versions of LingMess (Otmazgin et al., 2023), each trained on one of the training datasets. These are currently the best performing decoder and encoder-based models on the OntoNotes test set. For all models, hyperparameters are presented in Appendix B.

LinkAppend. Bohnet et al. (2023) proposed a method for formulating CR modeling as a sequence-to-sequence task where the objective is to predict links between mentions—forming a new coreference cluster—or to append mentions to existing clusters. For this, they trained a 13B parameter mT5 language model (Xue et al., 2021), with speaker and genre metadata, which achieved state-of-the-art performance on OntoNotes. We use the publicly released weights of the best performing trained model which we refer to as LinkAppend.

LingMess. Otmazgin et al. (2022) proposed an encoder-based CR model architecture and training procedure based on the mention ranking paradigm. They used this training recipe to train a state-of-the-art model, named LingMess, on the OntoNotes dataset. This model was initialized with a Longformer-large model as the pretrained encoder (Beltagy et al., 2020). To evaluate LingMess, we train a model on each of the three training sets using the default hyperparameters and Longformer-

large encoder. All training runs take at most 24 hours on a single 40GB A100 GPU.

5 Results

To examine how model performance varies across different types of coreference and different datasets, we conduct an extensive disaggregated evaluation. Summary results are presented in Table 4 for models evaluated on OntoNotes and OntoGUM, and Table 5 for all other test sets (exact AGG and TGG values are presented in Appendix D).

We illustrate how model failures correlate with differences between training and test sets in the way coreference is both defined and operationalized. In some of the datasets we studied (e.g., PreCo), the coreference types we consider in our study are fairly prevalent and represent up to 34% of all mentions annotated as coreferring (Table 5).

We observe significant performance variations for types of coreference that are defined differently across datasets. Consider the coreference types highlighted in Table 4. There are, for instance, significant performance variations (e.g., a performance drop of over 40 percentage points) for both generic and copular mentions between the in-domain—i.e., trained on a dataset constructed using the same definition of coreference as the test dataset—and out-of-domain—i.e., trained on datasets constructed using a different definition of coreference—LingMess models. While this variation might be explained by a number of factors, both PreCo and Phrase Detectives (used to train the out-of-domain models) differ from OntoNotes

Eval.	LingMess			# Ments.
	ON	PC	PD	
PreCo (C)	64.5	77.1	50.8	
PreCo (B^3)	64.2	76.3	47.3	25,983
Nested	11.6	55.9	0.5	1,760
ON Generic	12.0	70.5	50.9	3,520
Compound	5.0	77.1	0.0	2,163
Copular	0.8	63.7	0.0	1,410
Phrase Det. (C)	70.8	60.9	44.3	
Phrase Det. (B^3)	69.8	56.9	36.8	3,259
Nested	27.9	3.4	4.0	74
ON Generic	8.3	42.9	44.4	296
Compound	8.1	26.6	0.0	127
Copular	0.0	0.7	0.0	10
LitBank (C)	65.5	50.9	36.6	
LitBank (B^3)	66.3	47.0	32.9	2,267
Nested	24.3	7.3	0.0	39
ON Generic	5.7	16.5	13.2	91
Compound	0.0	0.0	0.0	11
Copular	28.6	6.7	0.0	10
ARRAU (C)	59.4	53.5	38.1	
ARRAU (B^3)	57.8	51.2	33.2	6,132
Nested	7.0	1.3	1.1	284
ON Generic	7.3	48.4	40.2	924
Compound	36.2	55.3	0.0	545
Copular	13.1	2.3	13.9	32
MMC (C)	61.4	55.6	45.2	
MMC (B^3)	57.3	51.0	40.5	5,338
Nested	10.0	11.0	0.0	140
ON Generic	16.6	53.0	34.8	393
Compound	0.0	10.3	100.0	7
Copular	1.2	55.1	1.2	636

Table 5: Disaggregated F1 scores using the same setup as Table 4, as evaluated on additional test sets. F1 scores are highlighted when the difference relative to the in-domain model is significantly larger than the same difference calculated for all mentions. C indicates CoNLL F1 score.

in that these types of mentions are considered part of the definition of coreference in both of these datasets. As a result, for instance, while only a few copular structures are annotated as coreferring in OntoNotes, LingMess$_{PC}$ predicts 1,158 coreferring mentions in a copular structure; *e.g.,* this example of a copular structure correctly predicted (according to the definitions used by PreCo and Phrase Detectives) as coreferring by LingMess$_{PC}$:

> . . . [One of the two honorable guests in the studio]$_1$ is [Professor Zhou Hanhug]$_1$ . . .

Consequently, both estimations of performance as well as measurements of model generalization will necessarily depend on which definition is used.

We also observe performance variations when there are operationalization differences across datasets constructed using the same definition of coreference. Consider the case of OntoGUM in Table 4, a dataset constructed using the same definition of coreference as OntoNotes. This allows us to examine the impact of potential differences in how the same definition of coreference was operationalized across datasets. Notably, models trained on

OntoNotes have poor performance on OntoGUM nested mentions despite these datasets also using the same annotation guidelines. While it is hard to fully disentangle why these differences exists, we conjecture they might be in part due to how these guidelines were interpreted by annotators: for instance, the OntoGUM paper describes removing i-within-i coreference relations; upon manual inspection of the cases where a model predicts a nested mention in the OntoGUM dataset that is not in the OntoGUM annotations, many of these purported failures appear to be attributable to this removal as all models predict certain i-within-i constructions. Consider the following nested coreference relations correctly predicted by LinkAppend, but not annotated in OntoGUM:

> Open-air markets, bookstores, and a [a Bart station all [its]$_1$ own]$_1$ make Rockridge another of many pleasant stops . . .

> So there's nothing too special about [the bed [itself]$_1$]$_1$, but it does have a very important function . . .

Such differences in how the same type of coreference are operationalized by dataset builders can further distort measurements of both model performance and model generalization.

6 Discussion

One might find it intuitive that measurements of CR model performance and generalization are likely to be impacted by differences in how coreference is defined and operationalized across datasets; and yet, existing multi-dataset evaluations do not explicitly considered how these differences might impact the performance estimates and the conclusions we can draw about models ability to generalize.

Existing evaluations of model generalization across datasets mostly report accuracy using aggregate metrics (Moosavi and Strube, 2018; Bamman et al., 2020; Toshniwal et al., 2021). Such aggregations, however, can obscure certain systematic errors (Hutchinson et al., 2022). Understanding of what exactly is being measured requires more explicit consideration of the measurement process. For this reason, we believe measurement modeling provides a useful framework for analyzing measurements in CR.

Our findings also highlight how relying on aggregated evaluations can obscure why models fail to generalize across datasets. We find this to be especially true in the case of coreference which involves

multiple sub-phenomena that are inconsistently annotated across datasets. If we do not resolve inconsistencies and errors in existing datasets, it will remain difficult to accurately interpret performance evaluation results based on these datasets.

6.1 Relative Ranking of Model Accuracy

A question one might ask is whether we can still make valid conclusions regarding the relative ranking of models' accuracy despite the absolute scores suffering from the issues with measurement validity that we have shown. Ultimately, however, ranking models based on empirical indicators is still a form of measurement and will therefore be subject to the same considerations that we have raised; *e.g.,* Is there a consistent definition of the theoretical concept being measured? Are the measurements valid in that they encompass only those variables we are intending to measure?

Consider the following empirical example, comparing LinkAppend and LingMess$_{ON}$: on 4 of the 5 "out-of-domain" test sets from Table 5, LinkAppend has a higher CoNLL F1 score than LingMess$_{ON}$. On the LitBank test set, however, LingMess$_{ON}$ has a 65.5 CoNLL F1 score versus 64.3 for LinkAppend. A possible explanation for this relative ranking is that LingMess$_{ON}$ has a lower recall across datasets (e.g., predicts fewer mentions in general), and LitBank is only annotated for a restricted set of mention types. Therefore, underfitting on the OntoNotes training set might lead to a better score on LitBank due to differences in how coreference is defined in these datasets rather than the ability of a model to generalize to unseen coreference links. Comparing the relative ranking of models will thus have limited "convergent validity" (*do measurements correlate with other measurements of the same construct?*) (Jacobs and Wallach, 2021) because the ranking is not consistent across datasets.

7 Conclusion

We propose viewing the measurement of coreference, CR model performance, and CR model generalization from the perspective of measurement modeling. To do so requires clearly distinguishing the theoretical construct intended to be measured from its measurement via a measurement model. This allows one to more clearly consider what is being measured, and possible concerns with the validity of these measurements.

Through a disaggregated, out-of-domain evaluation of CR models, we have shown that models appear systematically limited in their ability to generalize to certain types of coreference that differ in how they are defined or operationalized across datasets. We showed that models ostensibly display limited generalization for types of coreference even when evaluated on datasets intended to use the same annotation guidelines.

If the research goal is to develop CR models that generalize a consistent conceptualization across datasets, iteratively improving in-domain performance will have limited impact, as models' ability to generalize will inherently depend on operationalization differences between datasets and evaluation schemes.[5]

Measurement Modeling as a Blueprint. Finally, our work provides a blueprint for evaluating out-of-domain performance in a way that accounts for known inconsistencies in how coreference is defined and operationalized across datasets. In particular, our work highlights how explicitly considering the definition and operationalization of the construct being measured, as well as disaggregating performance results, can be used to better understand what we can learn from multi-dataset evaluations about models' ability to generalize. This is a necessary step in meaningfully evaluating models of contested constructs across multiple datasets.

7.1 Limitations and Future Work

Our empirical study only establishes correlations between model performance and types of CR when assessed across and within datasets. While there are certainly many more types of coreference that could be considered, by focusing on a select few types, we are able to study their interaction with dataset operationalization in greater detail.

Future work could explore other dimensions of operationalization such as: 1) the data processing (e.g., tokenization method); 2) the CR task format; 3) and the annotation procedure, including noise and ambiguity in annotation. Ultimately, a more holistic understanding of the way that coreference is operationalized can provide better insights into model accuracy and guide future modeling decisions. A promising direction might be to define such differences using grounded theory (Strauss and Corbin, 1997) similar to the method used by

---

[5]Our code is available at `https://github.com/ianporada/coref-data`

Robertson et al. (2021).

Furthermore, while one might intuitively expect that simply taking the intersection of all coreference types annotated in existing corpora can be used to measure model accuracy, the inclusion of certain phenomenon is not always discrete or easily determinable. In addition, taking an intersection of existing annotations does not solve of the problem of needing to clearly define the construct being measured. Rather, focusing on datasets with minimal differences is a possible direction for untangling the impact of any individual factor.

8 Ethics Statement

We have focused on evaluation datasets of CR, but we did not quantify possible downstream implications of these findings. Different types of CR might also be more prevalent in certain types of corpora that might be about or written by minoritized groups, or that might cover sensitive topics. CR models have been shown to exhibit biases, inferring coreferences disparately for distinct social groups (Webster et al., 2018; Kocijan et al., 2021; Hossain et al., 2023). Similarly, CR and other NLP datasets are also known to contain biased, stereotypical, or in other way problematic context (Cao and Daumé III, 2021; Selvam et al., 2022).

Acknowledgements

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

BBN Technologies. 2007. Co-reference guidelines for english ontonotes. Version 7.0. Guidelines accompanying OntoNotes 5.0 data release.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Anol Bhattacherjee. 2012. *Social science research: Principles, methods, and practices*. USA.

Thomas R. Black. 1999. Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics. *Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics.*, pages xiv, 751–xiv, 751. Place: Thousand Oaks, CA Publisher: Sage Publications Ltd.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

K. A. Bollen. 2001. Measurement models. In Neil J. Smelser and Paul B. Baltes, editors, *Indicator: Methodology*, International Encyclopedia of the Social & Behavioral Sciences, chapter 2, pages 7282–7287. Pergamon, Oxford.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661.

Haixia Chai, Nafise Sadat Moosavi, Iryna Gurevych, and Michael Strube. 2022. Evaluating coreference resolvers on community-based question answering: From rule-based to state of the art. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 61–73, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Haixia Chai and Michael Strube. 2023. Investigating multilingual coreference resolution by universal annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10010–10024, Singapore. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Nancy Chinchor. 1992. The statistical significance of the MUC-4 results. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

H. H. Clark. 1977. *Bridging*, pages 411–420. Cambridge University Press.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Ellen Edes. 1968. Output conditions on anaphoric expressions with split antecedents. *unpublished paper, Harvard University*.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Nupoor Gandhi, Anjalie Field, and Emma Strubell. 2022. Mention annotations alone enable efficient domain adaptation for coreference resolution.

Ralph Grishman and Beth Sundheim. 1996. Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*, pages 413–422, Vienna, Virginia, USA. Association for Computational Linguistics.

Graeme Hirst. 1981. *Anaphora*, pages 4–32. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1859–1876, New York, NY, USA. Association for Computing Machinery.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Robert Neal Kantor. 1977. *The management and comprehension of discourse connection by pronouns in English*. The Ohio State University.

R.B. Kline. 2011. *Principles and Practice of Structural Equation Modeling*. Methodology in the social sciences. Guilford Publications.

Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. The gap on gap: Tackling the problem of differing data distributions in bias-measuring datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13180–13188.

Sandra Kübler and Desislava Zhekova. 2011. Singletons and coreference resolution evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. ParCorFull2.0: a parallel corpus annotated with full coreference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 805–813, Marseille, France. European Language Resources Association.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486.

Jing Lu and Vincent Ng. 2020. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2018. Using linguistic features to improve the generalization capability of neural coreference resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Brussels, Belgium. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdenek Žabokrtskỳ, and Daniel Zeman. 2021. Coreference meets universal dependencies–a pilot experiment on harmonizing coreference datasets for 11 languages. *ÚFAL MFF UK, Praha, Czechia*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. Lingmess: Linguistically informed multi expert scorers for coreference resolution.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, Denver, Colorado. Association for Computational Linguistics.

M. Poesio, F. Bruneseaux, and L. Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Towards Standards and Tools for Discourse Tagging*.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9:561–587.

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 79–90, Barcelona, Spain (online). Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "i can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Ina Roesiger, 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.

Nikil Roashan Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2022. The tail wagging the dog: Dataset construction biases of social bias benchmarks. *arXiv preprint arXiv:2210.10040*.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.

Sanjay Subramanian and Dan Roth. 2019. Improving generalization in coreference resolution via adversarial training. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. Deep cross-lingual coreference resolution for less-resourced languages: The case of Basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 35–41, Minneapolis, USA. Association for Computational Linguistics.

Olga Uryupina. 2008. Error analysis for learning-based coreference resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3):333–353.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive processes*, 6(2):107–135.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Ralph Weischedel et al. 2013. Ontonotes release 5.0.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor Wai-Hung Tsang, Kian Ming Adam Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 744–753, Jeju Island, Korea. Association for Computational Linguistics.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2023a. The universal anaphora scorer 2.0. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 183–194, Nancy, France. Association for Computational Linguistics.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2023b. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and Wikipedia texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 767–781, Dubrovnik, Croatia. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk,

Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612.

Amir Zeldes. 2022. Opinion piece: Can we fix the scope for coreference? *Dialogue & Discourse*, 13(1):41–62.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 92–101, San Diego, California. Association for Computational Linguistics.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 21–29, Gothenburg, Sweden. Association for Computational Linguistics.

Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, and Benjamin Van Durme. 2023. Multilingual coreference resolution in multiparty dialogue. *Transactions of the Association for Computational Linguistics*, 11:922–940.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

Heike Zinsmeister and Stefanie Dipper. 2010. Towards a standard for annotating abstract anaphora. In *Workshop on Language Resource and Language Technology Standards at the International Conference on Language Resources and Evaluation*.

A   Additional Dataset Details

A.1   Dataset Selection and Scope

A.1.1   Dataset Selection

Our selection of datasets is based on the multi-dataset evaluations of Toshniwal et al. (2021), Xia and Van Durme (2021), Zhu et al. (2021), and Žabokrtský et al. (2023).

Of the datasets evaluated by Xia and Van Durme (2021) (X&V), we select the same large-scale training sets, OntoNotes and PreCo, as well as the two largest English-language test sets, LitBank and AR-RAU. The collection of Toshniwal et al. (2021) is mostly a subset of X&V, with the exception of the Friends dataset (Zhou and Choi, 2018). We use the more recent MMC dataset (Zheng et al., 2023) which encompasses the Friends corpus.

Žabokrtský et al. (2023)'s evaluation includes two English-language datasets: GUM (Zeldes, 2017) and English ParCorFull (Lapshinova-Koltunski et al., 2022). Of these we select GUM, as ParCorFull consists of only 19 documents. For GUM, we specifically use the OntoGUM formatted version of the dataset from Zhu et al. (2021) which allows us to consider the case where two datasets are intended to share definitions of the same theoretical construct, but differ in aspects related to their operationalization.

Furthermore, we include the recently released, large-scale Phrase Detectives 3.0 dataset (Yu et al., 2023b) as a training set.

A.1.2   Dataset Scope

While *coreference* is most often used to refer to *identity coreference*, a relationship between linguistic expressions that refer to discourse entities with the same identity (Nedoluzhko et al., 2021), there are many related phenomena which are sometimes also referred to as coreference—and may be considered instances of identity coreference in certain contexts—such as bridging anaphora (Clark, 1977; Roesiger et al., 2018), discourse deixis (Webber, 1991; Zinsmeister and Dipper, 2010), event coreference (sometimes considered a subtype of discourse deixis) (Cybulska and Vossen, 2014; Lu and Ng, 2018), and split-antecedent anaphora (Edes, 1968). We focus on datasets of identity coreference and consider these additional types only insofar as they are included within the annotations of a dataset being studied.

Furthermore, as in Xia and Van Durme (2021), we exclude discontinuous mention spans (Yu et al.,

Dataset	Generic Only	VPs	Appositives	Copulae
OntoNotes	✗	✓	✗*	✗
PreCo	✓	✗	✓	✓
Phrase Det.	✓	✗	✗*	✗*

Table 6: Differences in definitions of *coreference* as a theoretical construct (*annotated in the dataset, but not considered coreference).

2023a) from our evaluations. Discontinuous spans are incompatible with most existing CR models which rely on the assumption that mentions correspond to a continuous span.

The phenomenon of *singletons* is also sometimes considered related to identity coreference, although definitions of singletons differ. Kübler and Zhekova (2011) describe singletons as "a cover term for mentions that are never coreferent, such as in *in general* or *on the contrary*, and mentions that are potentially coreferent but occur only once in a document." However, another common definition of a singleton is a linguistic expression that refers to a discourse entity which is referenced only once (Zeldes, 2022). The existence of singletons is one aspect of the definition of coreference that has been considered in recent multi-dataset evaluations (Toshniwal et al., 2021). We focus our evaluation on non-singleton mentions, a common practice for multi-dataset evaluations (Žabokrtský et al., 2022).

A.1.3 Dataset Formatting

For PreCo, we use the last 500 documents of the training set as a validation split. For Phrase Detectives we randomly select 45 documents from the training set as a validation split. For Litbank we use "split_0" from the official repository. For OntoGUM we use the official splits and include Reddit data. For ARRAU, we use the existing splits for the WSJ data and randomly split the remaining data into train/validation/test splits of size 80/10/10 percent, respectively.

A.2 Detailed Constructs

Differences in constructs are presented in Table 6. We provide additional details below.

OntoNotes. *Identical coreference* is defined in the OntoNotes annotation guidelines to be "names, nominal mentions, pronominal mentions, and verbal mentions of the same entity, concept, or event" (BBN Technologies, 2007). This construct is distinguished from appositives ("immediately-adjacent noun phrases, separated only by a comma, colon, dash, or parenthesis") and copular structures (a subject and predicate linked by a copula). Fur-

thermore, identical coreference is defined to "not include entities that are only mentioned as generic, underspecified or abstract."

PreCo. The PreCo authors establish a conceptualization of coreference mostly by contrasting what is considered to be coreferring in PreCo with the OntoNotes definition. As the PreCo annotation guidelines are not public, we can only assess the construct intended to be measured based on the corresponding paper and data. The PreCo authors note that they follow most of the conceptualization of coreference used in OntoNotes with the exception that they exclude verb phrases and explicitly include generic mentions and mentions in appositive and copular structures which may refer to the same entity. The PreCo dataset is additionally annotated for singleton mentions, although no definition is given for the concept of singleton.

Phrase Detectives. In the case of Phrase Detectives, coreference is considered a relation strictly between noun phrases. Predication, including apposition and copular structures, is distinguished from coreference and annotated separately. Generic only mentions can be annotated as coreferring.

B Model Hyperparameters

For all models we use the default hyperparameters, except that we use bf16 precision for training of LingMess models in order to decrease the total training time required.

C Statistical Significance

We perform statistical significance tests using the same procedure as the MUC evaluation as described by Chinchor (1992). That is, we conduct a permutation test with 10,000 permutations and an α of 0.1. We use this significance test to highlight F1 scores where the difference relative to the in-domain model is significantly larger than the same difference calculated for all mentions.

D Additional Results

D.1 Additional Types

While we do not focus on these types in the body of the paper, the noted differences in the annotation of verb phrases (VPs) and apposition can also be heuristically approximated for which we observe trends similar to the other types.

LingMess$_{PC}$ and LingMess$_{PD}$ never predict VPs as coreferring in the OntoNotes test set (where a

VP is taken to be a mention that consists of only V* POS tags) compared with 15.2 / 100.0 / 26.3 disaggregated R / P / F1 for LingMess$_{ON}$.

Appositives are annotated in OntoNotes and Phrase Detectives, but not considered identical coreference. If we approximate apposition as two coreferring mentions in the same sentence that are either adjacent or separated only by a single punctuation token, LingMess$_{ON}$ achieves 67.0 disaggregated F1 on appositives as compared to 13.3 and 25.0 for LingMess$_{PC}$ and LingMess$_{PD}$, respectively, again showing a systematic drop in performance.

D.2 Generalization Gaps

	LingMess$_{PC}$	LingMess$_{PD}$
OntoNotes	26.6	41.8
Nested	52.0	55.4
ON Generic	40.1	34.4
Compound	57.1	81.0
Copular	49.2	50.0
OntoGUM	14.3	22.7
Nested	0.0	0.0
ON Generic	13.1	14.4
Compound	26.3	47.5

Table 7: The AGG and TGG values calculated for Table 4. (I.e., the difference in the F1 score of the out-of-domain model as compared to the in-domain model.)

	LingMess$_{PC}$	LingMess$_{PD}$
PreCo	12.1	16.9
Nested	44.3	11.1
ON Generic	58.5	38.9
Compound	72.1	5.0
Copular	62.9	0.8
Phrase Det.	12.9	33.0
Nested	24.5	23.9
ON Generic	34.6	36.1
Compound	18.5	8.1
Copular	0.7	0.0
LitBank	19.3	33.4
Nested	17	24.3
ON Generic	10.8	7.5
Compound	0.0	0.0
Copular	21.9	28.6
ARRAU	6.6	24.6
Nested	5.7	5.9
ON Generic	41.1	32.9
Compound	19.1	36.2
Copular	10.8	0.8
MMC	6.3	16.8
Nested	1.0	10.0
ON Generic	36.4	18.2
Compound	10.3	100.0
Copular	53.9	0.0

Table 8: The AGG and TGG values calculated for Table 5. (I.e., the difference in the F1 score of the out-of-domain model as compared to the in-domain model.)