ezCoref: Towards Unifying Annotation Guidelines for Coreference Resolution

Ankita Gupta[♠] **Marzena Karpinska**[♠] **Wenlong Zhao**[♠] **Kalpesh Krishna**[♠]
Jack Merullo[♢] **Luke Yeh**[♡] **Mohit Iyyer**[♠] **Brendan O'Connor**[♠]

[♠]University of Massachusetts Amherst, [♢]Brown University, [♡]Google
{ankitagupta,mkarpinska,wenlongzhao,kalpesh,miyyer,brenocon}@cs.umass.edu
john_merullo@brown.edu, lukeyeh@google.com

Abstract

Large-scale, high-quality corpora are critical for advancing research in coreference resolution. However, existing datasets vary in their definition of coreferences and have been collected via complex and lengthy guidelines that are curated for linguistic experts. These concerns have sparked a growing interest among researchers to curate a unified set of guidelines suitable for annotators with various backgrounds. In this work, we develop a crowdsourcing-friendly coreference annotation methodology, ezCoref, consisting of an annotation tool and an interactive tutorial. We use ezCoref to re-annotate 240 passages from seven existing English coreference datasets (spanning fiction, news, and multiple other domains) while teaching annotators only cases that are treated similarly across these datasets.[1] Surprisingly, we find that reasonable quality annotations were already achievable (>90% agreement between crowd and experts) even without extensive training. On carefully analyzing the remaining disagreements, we identify the presence of linguistic cases that our annotators unanimously agree upon but lack unified treatments (e.g., generic pronouns, appositives) in existing datasets. We propose the research community should revisit these phenomena when curating future unified annotation guidelines.

1 Introduction

Coreference resolution is the task of identifying and clustering together all textual expressions (*mentions*) that refer to the same discourse entity in a given document. Impressive progress has been made in developing coreference systems (Lee et al., 2017; Moosavi and Strube, 2018; Joshi et al., 2020), enabled by datasets annotated by experts (Hovy et al., 2006; Bamman et al., 2020; Uryupina et al., 2019) and crowdsourcing (Chamberlain et al., 2016). However, these datasets vary widely in their definitions of coreference (expressed via annotation guidelines), resulting in inconsistent annotations both within and across domains and languages. For instance, as shown in Figure 1, while ARRAU (Uryupina et al., 2019) treats generic pronouns as non-referring, OntoNotes (Hovy et al., 2006) chooses not to mark them at all.

OntoNotes: Maybe we need a [CIA] version of the Miranda warning: You have the right to conceal your coup intentions, because we may rat on you.

ARRAU: Maybe [we]e1 need [a [CIA] version of [the Miranda warning]]: [You]e4 have [the right to conceal [[your]e5 [coup] intentions]], because [we]e6 may rat on [you]e7.

Crowd (this work): Maybe [we]e1 need [a [CIA] version of [the [Miranda] warning]]: [You]e3 have [the right] to conceal [[your]e3 coup intentions], because [we]e1may rat on [you]e3.

Figure 1: We visualize a common sentence from news domain annotated by two expert-curated datasets, OntoNotes (Hovy et al., 2006) and ARRAU (Uryupina et al., 2019), along with the crowd annotations collected via our ezCoref platform. OntoNotes does not mark generic pronouns. ARRAU does not consider them as coreferent and annotates them using a special relation "undef-reference" (markables with vague interpretations). On the contrary, our crowdworkers assign all mentions of the generic pronoun "you" to the same coreference chain. The situation is also similar for the generic "we."

It is thus unclear which guidelines one should employ when collecting coreference annotations in a new domain or language. Traditionally, existing guidelines have leaned towards lengthy explanations of complex linguistic concepts, such as those in the OntoNotes guidelines (Weischedel et al., 2012), which detail what should and should not be coreferent (e.g., how to deal with head-sharing noun phrases, premodifiers, and generic mentions). As a result, coreference datasets have traditionally been annotated by linguists (experts) already familiar with such concepts, which makes the process expensive and time-consuming. Crowd-

[1]Our platform's code and collected data is available at https://github.com/gnkitaa/ezCoref

Findings of the Association for Computational Linguistics: EACL 2023, pages 312–330
May 2-6, 2023

sourced coreference data collection has the potential to be significantly cheaper and faster; however, teaching an exhaustive set of linguistic guidelines to non-expert crowd workers remains a formidable challenge. As a result, there has been a growing interest among researchers in curating a unified set of guidelines (Poesio et al., 2021) suitable for annotators with various backgrounds.

More recently, games-with-a-purpose (GWAPs) (von Ahn, 2006; Poesio et al., 2013) were proposed to aid crowdsourcing of large coreference datasets (e.g., Chamberlain et al., 2016; Yu et al., 2022). While GWAPs make it enjoyable for crowdworkers to learn complex guidelines and perform annotations using them (Madge et al., 2019b), they also require significant effort to attract and maintain workers. For instance, Phrase Detectives Corpus 1.0 was collected over a span of six years (Chamberlain et al., 2016; Poesio et al., 2013; Yu et al., 2022), which motivates us to instead study coreference collection on more efficient payment-based platforms.

Specifically, our work investigates the quality of crowdsourced coreference annotations when annotators are taught only simple coreference cases that are treated uniformly across existing datasets (e.g., pronouns). By providing only these simple cases, we are able to teach the annotators the concept of coreference, while allowing them to freely interpret cases treated differently across the existing datasets. This setup allows us to identify cases where our annotators unanimously agree with each other but disagree with the expert, thus suggesting cases that should be revisited by the research community when curating future guidelines.

Our main contributions are:

- We develop a crowdsourcing-friendly coreference annotation methodology—`ezCoref`—which includes an intuitive, open-sourced annotation tool supported by a short crowd-oriented interactive tutorial.[2]
- We use `ezCoref` to re-annotate 240 passages from seven existing English coreference datasets on Amazon Mechanical Turk (AMT), and conduct a comparative analysis of crowd and expert annotations. We find that high-quality annotations are already achievable from non-experts without extensive training (>90% B3 (Bagga and Baldwin, 1998a) agreement between crowd and experts).
- We further qualitatively analyze remaining disagreements among crowd and expert annotations and identify linguistic cases that crowd unanimously marks as coreferent but lack unified treatment in existing datasets (e.g., generic pronouns as shown in Figure 1). Additionally, analyzing inter-annotator agreement among the crowd reveals that crowd exhibits higher agreement when annotating familiar texts (e.g., childhood stories or fiction) compared to texts rich in cataphoras or those requiring world knowledge. Finally, our qualitative analysis also provides an empirical evidence to support previous findings in literary studies (Szakolczai's (2016) analysis of Bleak House) and psychology (Orvell et al.'s (2020) claims about generic "you").

2 Related Work

Existing coreference datasets: Table 1 provides an overview of seven prominent coreference datasets, which differ widely in their annotator population, mention detection, and coreference guidelines.[3] Many datasets are annotated by experts heavily trained in linguistic standards, including ARRAU (Uryupina et al., 2019), LitBank (Bamman et al., 2020), GUM (Zeldes, 2017), and OntoNotes (Hovy et al., 2006). Due to its scale and quality, OntoNotes is likely the most widely used for NLP coreference research, including in two CoNLL shared tasks (Pradhan et al., 2011, 2012). QuizBowl (Guha et al., 2015) has been annotated by domain (but not linguistic) experts. Few coreference datasets exists which are annotated by non-experts, including those created by part-time non-native English speakers (PreCo; Chen et al., 2018), and gamified crowdsourcing without financial compensation (Phrase Detectives; Chamberlain et al., 2016; Yu et al., 2022).

Coreference annotation tools: Several coreference annotation tools have been developed (See Table A3 in Appendix for more details). However, these are difficult to port to a crowdsourced workflow, as they require users to install software on their local machine (Widlöcher and Mathet, 2012; Landragin et al., 2012; Kopeć, 2014; Mueller and Strube, 2001; Reiter, 2018), or have complicated

[2] Our tutorial received overwhelmingly positive feedback. One annotator commented that it was "*absolutely beautiful, intuitive, and helpful. Legitimately the best one I've ever seen in my 2 years on AMT! Awesome job.*" (Table A4 in Appendix)

[3] Many others exist too; for example, see Jonathan Kummerfeld's spreadsheet list (accessed Jan. 2022).

Dataset	Domains #(doc, ment, tok)	Annotators	Mention Detection	Mention Types		Coreference Links			
				Singletons	Entity Restrictions	Copulae	Appositives	Generics	Ambiguity
ARRAU (Uryupina et al., 2019)	Multiple (552, 99K, 350K)	Single Expert	Manual	Yes	None	Special Link	No Link	Yes	Explicit
OntoNotes (Hovy et al., 2006)	Multiple (1.6K, 94K, 950K)	Experts	Mixed	No	None	Special Link	Special Link	Only with Pronominals	None
LitBank (Bamman et al., 2020)	Single (100, 29K, 210K)	Experts	Manual	Yes	ACE (selected)	Special Link	Special Link	Only with Pronominals	None
GUM (Zeldes, 2017)	Multiple (25, 6K, 20K)	Experts (Linguistics Students)	Manual	Yes	None	Coref (Sub-Types)	Coref (Sub-Type)	Yes	None
QuizBowl (Guha et al., 2015)	Single (400, 9.4K, 50K)	Domain Experts	Manual & CRF*	Yes	Characters, Books, Authors*	Coref	Coref	If Applicable	None
PreCo** (Chen et al., 2018)	Multiple (38K, 3.58M, 12.5M)	Non-Expert, Non-Native	Manual**	Yes	None	Coref	Coref	Yes	None
Phrase Detectives (PD) (Chamberlain et al., 2016)	Multiple (542, 100K, 400K)	Crowd (gamified) + 2 Experts	Semi Automatic	Yes	None	Special Link	Special Link	Yes	Implicit
ezCoref Pilot Dataset (this work)	Multiple	Crowd (paid)	Fully Automatic	Yes	None	Annotator's Intuition	Annotator's Intuition	Annotator's Intuition	Implicit

Table 1: Summary of seven datasets analyzed in this work, which differ in domain, size, annotator qualifications, mention detection procedures, types of mentions, and types of links considered as coreferences between these mentions.*Allows other types of mention only when this mention is an answer to a question.**We interpret manual identification based on illustrations presented in the original publication (Chen et al., 2018). ***Inaccessible, see Footnote 8.

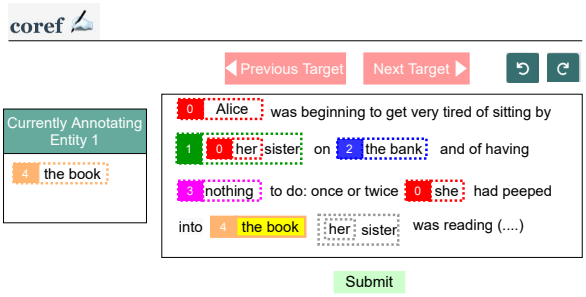

Figure 2: Part of the ezCoref interface (§3)

UI design with multiple drag and drop actions and/or multiple windows (Stenetorp et al., 2012; Widlöcher and Mathet, 2012; Landragin et al., 2012; Yimam et al., 2013; Girardi et al., 2014; Kopeć, 2014; Mueller and Strube, 2001; Oberle, 2018). Closest to ezCoref is CoRefi (Bornstein et al., 2020), a web-based coreference annotation tool that can be embedded into crowdsourcing websites. Subjectively, we found its user interface difficult to use (e.g., users have to memorize multiple key combinations). It also does not allow for nested spans, reducing its usability.

Crowdsourcing linguistic annotations: Several efforts have been made to crowdsource linguistic annotations (Snow et al., 2008; Callison-Burch, 2009; Howe, 2008; Lawson et al., 2010), including on payment-based microtasks via platforms like AMT and GWAPs (von Ahn, 2006). Many GWAPs (Poesio et al., 2013; Kicikoglu et al., 2019; Madge et al., 2019a; Fort et al., 2014) have been used in NLP to collect linguistic annotations including coreferences; with some broader platforms (Venhuizen et al., 2013; Madge et al., 2019b) aiming to gamify the entire text annotation pipeline. One solution to teaching crowd workers complex guidelines is to incorporate *learning by progression* (Kicikoglu et al., 2020; Madge et al., 2019b; Miller et al., 2019), where annotators start with simpler tasks and gradually move towards more complex problems, but this requires subjective judgments of task difficulty. In contrast to the payment-based microtask setting studied in this work, GWAPs are not open-sourced, need significant development, take longer to collect data, and require continuous efforts to maintain visibility (Poesio et al., 2013).

3 ezCoref: A Crowdsourced Coreference Annotation Platform

The ezCoref user experience consists of (1) a step-by-step interactive tutorial and (2) an annotation interface, which are part of a pipeline including automatic mention detection and AMT Integration.

Annotation structure: Two annotation approaches are prominent in the literature: (1) a local pairwise approach, annotators are shown a pair of mentions and asked whether they refer to the same entity (Hladká et al., 2009; Chamberlain et al., 2016; Li et al., 2020; Ravenscroft et al., 2021), which is time-consuming; or (2) a cluster-based approach (Reiter, 2018; Oberle, 2018; Bornstein et al., 2020), in which annotators group all mentions of the same entity into a single cluster. In ezCoref we use the latter approach, which can be faster but requires the UI to support more complex actions for creating and editing cluster structures.

Example	Phenomena Taught
[John] doesn't like [Fred], but [he] still invited [him]to [the party].	(1) personal pronouns (2) singletons
[This dog] likes to play [catch].[It]'s better than other [dogs] at [this game]. [[Its] owner] is really proud.	(1) possessive pronouns (2) semantically similar expression which are not coreferring (3) non-person entities (animals)
[Director [Mackenzie]] spent [last two years] working on a ["Young Adam"]. During [this time] [he] often had to make [compromises] but [the movie] turned out to exceed expectations.	(1) nested spans (2) non-person entities (time, item)
[The office] wasn't exactly small either. [I]'m sure that 50, or maybe even 60, [people] could easily fit [there].	(1) non-person entities (place)

Table 2: Simple coreference cases explained in tutorial.

User interface: We spent two years iteratively designing, implementing, and user testing the interface to make it as simple and crowdsourcing-friendly as possible (Figure 2).[4] Marked mentions are surrounded by color-coded frames with entity IDs. The currently selected mention ("the book"), is highlighted with a flashing yellow cursor-like box. The core annotation action is to select other mentions that corefer with the current mention, and then advance to a later unassigned mention; annotators can also re-assign a previously annotated mention to another cluster. Advanced users can exclusively use keyboard shortcuts, undo and redo actions were added to allow error correction. Finally, `ezCoref` provides a side panel showing mentions of the entity currently being annotated to spot mentions assigned to the wrong cluster.

Coreference tutorial: To teach crowdworkers the basic definition of coreference and familiarize them with the interface, we develop a tutorial (aimed to take $\sim$ 20 minutes) that introduces them to the mechanics of the annotation tool, and then trains them on simple cases of coreferences. These cases (e.g., personal/possessive pronouns or determinative phrases which corefer with their antecedents as shown in Table 2) are annotated similarly across all existing datasets and are unlikely to be disputed. The tutorial concludes with a quality control example to exclude poor quality annotators.[5] These training examples, feedback, and annotation guidelines can be easily customized using a simple JSON schema.

Annotation workflow: The annotators are presented with one passage (or "document") at a time (Figure 2), and all mentions have to be annotated before proceeding to the next passage. There is no limitation to the length or language of the passage.

[4]The interface is implemented in ReactJS.

[5]Examples of the tutorial interface and the quality control example are provided in Appendix.

In this work, we divide an initial document into a sequence of shorter passages of complete sentences, on average 175 tokens, as shorter passages minimize the need to scroll, reducing annotator effort. While this obviously cannot capture longer distance coreference,[6] a large portion of important coreference phenomena is local: within the OntoNotes written genres, for pronominal mentions, the closest antecedent is contained within the current or previous two sentences more than 95% of the time.

Automatic mention detection: As a first step to collect coreference annotations, we must identify mentions in the documents from each of the seven existing datasets; this process is done in a diverse array of ways (from manually to automatic) in prior work as shown in Table 1. We decided to automatically identify mentions to give all crowdworkers an identical set of mentions, which simplifies the annotation task and also allows us to easily compare and study their coreference annotations via inter-annotator agreement. Specifically, we implement a simple algorithm that yields a high average recall over all seven datasets.[7]

Our algorithm considers all noun phrases (including proper nouns, common nouns, and pronouns) as markables, extracting them using the Stanza dependency parser (version 1.3.0; Qi et al., 2020). We allow for nested mentions and proper noun premodifiers (e.g., [U.S.] in "U.S. policy"). We include all conjuncts with the entire coordinated noun phrase ([Mark], [Mary], as well as [Mark and Mary], are all considered mentions); details in Appendix A.3.

4 Using `ezCoref` to Re-annotate Existing Coreference Datasets

We deploy `ezCoref` on the AMT crowdsourcing platform to re-annotate 240 passages from seven existing datasets, covering seven unique domains. In total, we collect annotations for 12,200 mentions and 42,108 tokens. We compare our workers' an-

[6]We leave this for future work—for example, more sophisticated user interfaces to support longer documents, or merging coreference chains between short passages. As documents get progressively longer, such as book chapters or books, the task takes on aspects of cross-document coreference and entity linking (e.g. Bagga and Baldwin, 1998b; FitzGerald et al., 2021; Logan IV et al., 2021).

[7]We acknowledge that any algorithm can be used as long as its recall across all datasets is high, and ours is only one such algorithm. However, we do not conduct an ablation study to compare crowd annotations for mentions obtained from these potential algorithms as it would be prohibitively expensive. Furthermore, while advanced mention detection methods can improve annotation quality, our goal is not to collect the highest-quality coreference dataset, but to study annotator behavior when a common set of mentions is provided.

notations both quantitatively and qualitatively to each other and to existing expert annotations.

Datasets: We collect coreference annotations for the seven existing datasets described in Table 1: OntoNotes (Hovy et al., 2006), LitBank (Bamman et al., 2020), PreCo[8] (Chen et al., 2018), ARRAU (Uryupina et al., 2019), GUM (Zeldes, 2017), Phrase Detectives (Chamberlain et al., 2016), and QuizBowl (Guha et al., 2015). The sample covers seven domains: news, opinionated magazines, weblogs, fiction, biographies, Wikipedia articles, and trivia questions from Quiz Bowl. For each dataset with multiple domains, we manually select a broad range of domain(s) for re-annotation. From each domain in each dataset, we then select documents and divide them into shorter passages (on average 175 tokens each), creating 20 such passages per dataset. For datasets with multiple domains, we choose 20 such passages per domain (see Appendix A.1 for detail). Overall, we collect annotations for 240 passages with 5 annotations per passage to measure inter-annotator agreement.

Procedure: We first launch an annotation tutorial and recruit the annotators on the AMT platform.[9] At the end of the tutorial, each annotator is asked to annotate a short passage (around 150 words). Only annotators with a B3 score (Bagga and Baldwin, 1998a) of 0.90 or higher are then invited to participate in the annotation task.

Training Annotators with Simplified Guidelines using `ezCoref`: As the goal of our study is to understand what crowdworkers perceive as coreference, we train our annotators with simple guidelines. We carefully draft our training examples to include only cases which are considered as coreference by all the existing datasets. The objective is to teach crowdworkers the broad definition of coreference while leaving space for different interpretations of ambiguous cases or those resolved differently across the existing datasets. Note that a comparable experiment with more complex guidelines is infeasible since it is unclear which guidelines to choose, and also providing complex linguistic guidelines to crowdworkers remains an open challenge. Overall, `ezCoref` is aimed to minimize both researcher and annotator effort for new coreference data collection, compared to prior work (Figure 3).

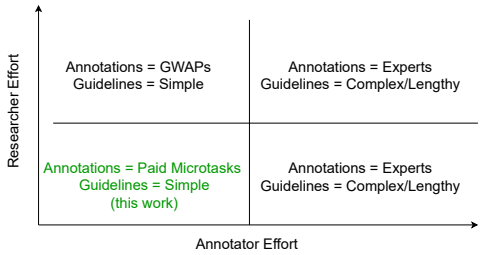

Figure 3: Existing expert annotated datasets entail high annotator effort (e.g., OntoNotes, ARRAU). Existing crowdsourced coreference datasets (e.g., Phrase Detectives) entail significant researcher effort. In this work, we explore the minimum effort scenario for both annotators (by providing them simplified guidelines) and researchers (by open-sourcing `ezCoref`).

Worker details: Overall, 73 annotators (including 44 males, 20 females, and one non-binary person)[10] completed the tutorial task, which took 19.4 minutes on average (sd=11.2 minutes). They were aged between 21 and 69 years (mean=38.9, sd=11.3) and identified themselves as native English speakers. Most of the annotators had at least a college degree (47 vs 18). 89.0% of annotators, who did the tutorial, received a B3 score of 0.90 or higher for the final screening example, and were invited to the annotation task. 50.7% of the invited annotators returned to participate in the main annotation task, and 29.2% of them annotated five or more passages. Annotation of one passage took, on average, 4.15 minutes, a rate of 2530 tokens per hour. The total cost of the tutorial was $460.70 ($4.50 per tutorial). We paid $1 per passage for the main annotation task, resulting in a total cost of $1440.[11]

5 Analysis

In this section, we perform quantitative and qualitative analyses of our crowdsourced coreference annotations. First, we evaluate the performance

[8]The PreCo dataset is interestingly large but seems difficult to access. In November 2018 and October 2021 we filled out the data request form at the URL provided by the paper, and attempted to contact the PreCo official email directly, but did not receive a response. To enable a precise research comparison, we scraped all documents from PreCo's public demo in November 2018 (no longer available as of 2021); its statistics match their paper and our experiments use this version of the data. PreCo further suffers from data curation issues (Gebru et al., 2018; Jo and Gebru, 2020); it uses text from English reading comprehension tests collected from several websites, but the original document sources and copyright statuses are undocumented. When reading through PreCo documents, we found many domains including opinion, fiction, biographies, and news (Table A1 in Appendix); we use our manual categories for domain analysis.

[9]We allow only workers with a >= 99% approval rate and at least 10,000 approved tasks who are from the US, Canada, Australia, New Zealand, or the UK.

[10]We did not collect demographic data for the remaining eight individuals, from an earlier pilot experiment.

[11]All reported costs include 20% AMT fee.

of our mention detection algorithm, comparing it to gold mentions across seven datasets. Next, we measure the quality of our annotations and their agreement with other datasets. Finally, we discuss interesting qualitative results.

5.1 Mention Detector Evaluation

Datasets differ in the way they define their mention boundaries and thus the boundaries for the same mention may differ. To fairly compare our mentions with the gold standards, we employ a headword-based comparison. We find the head of the given phrase by identifying, in the dependency tree, the most-shared ancestor of all tokens within the given mention. Two mentions are considered same if their respective headwords match.

Table 3 compares our mention detector to the gold mentions in existing datasets. Our method obtains high recall across most datasets (>0.90), which shows that most of the mentions annotated in existing datasets are correctly identified and allows a direct comparison of crowd annotations with expert annotations. It has the lowest recall with ARRAU (0.84) and PreCo (0.88), which is to be expected as ARRAU marks all referring premodifiers (identified manually) and PreCo allows common noun modifiers, while we identify only the premodifiers which are proper nouns.[12]

For most datasets, the precision is >0.80, suggesting that the algorithm identifies most of the relevant mentions. We observe a substantially lower score for OntoNotes, LitBank, and QuizBowl as these datasets restrict their mention types to limited entities (refer to Table 1). However, this does not limit our analysis. In fact, an algorithm with high precision on LitBank or OntoNotes would miss a huge percentage of relevant mentions and entities on other datasets (constraining our analysis) and when annotating new texts and domains. Furthermore, our algorithm identifies more mentions than in the original datasets, which in the best case allows us to discover new entities and, in the worst case, may result in more singletons Finally, the mention density (number of mentions per token) from our detector remains roughly consistent across all datasets when using our method, allowing us to fairly compare statistics (e.g., agreement rates) across datasets.

[12]We made this decision as identifying automatically all premodifiers would result in many singletons and lead to more arduous annotation effort.

Dataset	Recall	Precision	Mentions / Tokens	
			Gold	This Work
OntoNotes	0.957	0.376	0.112	0.286
LitBank	0.962	0.415	0.121	0.280
QuizBowl	0.956	0.543	0.188	0.318
PD (Gold)	0.953	0.803	0.259	0.273
PD (Silver)	0.938	0.791	0.265	0.274
GUM	0.906	0.848	0.269	0.287
PreCo	0.881	0.883	0.287	0.287
ARRAU	0.840	0.870	0.289	0.279

Table 3: Comparison of mentions identified by our mention detection algorithm with the gold mentions annotated in the respective datasets. We use head-word based comparison to compare mentions of different lengths. Our method obtains high recall across most datasets and the mention-density using our mention-detector remains roughly consistent across datasets, allowing us to do fair analysis (e.g., agreement) across datasets.

5.2 Agreement with Existing Datasets

How well do annotations from `ezCoref` agree with annotations from existing datasets?

Aggregating annotations: To compare crowdsourced annotations with gold annotations, we first require an aggregation method that can combine annotations from multiple crowdworkers to infer coreference clusters. We use a simple aggregation method that determines whether a pair of mentions is coreferent by counting the number of annotators who marked the two mentions in the same cluster.[13] Two mentions are considered as coreferent when the number of annotators linking them together is greater than a threshold (τ). After inferring these pairs of mentions, we construct an undirected graph where nodes are mentions and edges represent coreference links. Finally, we find connected components in the graph to obtain coreference clusters.[14] We compare aggregated annotations from `ezCoref` with gold annotations across the seven datasets using B3 scores (precision, recall, and F1),[15] as illustrated in Figure 4.

High agreement with OntoNotes, GUM, LitBank, ARRAU: Our annotators achieve the high-

[13]Future data collection efforts interested in creating large resources can utilize more advanced aggregation methods (Poesio et al., 2019).

[14]This method resolves to majority voting-based aggregation when the τ is set so that more than half of annotators should agree. For $\tau = N$, this method is very conservative, adding a link between two mentions only when all annotators agree unanimously. Conversely, for $\tau = 1$, only a single vote is required to add a link between two mentions.

[15]For a mention in a given document, B3 recall is the fraction of mentions that are correctly predicted by the system as coreferent with it out of all mentions that are actually coreferent with it. B3 precision is the fraction of mentions that are correctly predicted by the system as coreferent with it out of all system-predicted mentions.

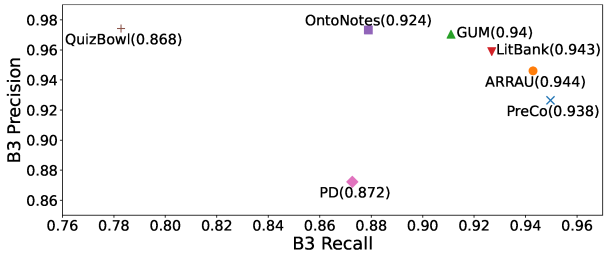

Figure 4: Agreement with gold annotations across datasets. B3 (F1) scores shown in parentheses are computed with singletons included.

est precision with OntoNotes (Figure 4), suggesting that most of the entities identified by crowdworkers are correct for this dataset. In terms of F1 scores, the datasets which are closest to crowd annotations are GUM, LitBank, and ARRAU, all of which are annotated by experts. This result shows that high-quality annotations can be obtained from non-experts using `ezCoref` without extensive training. We further conducted a qualitative analysis of high agreement cases for each dataset. Overall, we observe that non-experts agree with experts on chains containing pronouns and named entities. However, non-experts also mark noun phrases in appositive constructions as coreferent, consistent with GUM guidelines. Finally, non-experts also assign generic mentions to the same coreference chain, consistent with their treatment by GUM and ARRAU, and leads to higher agreement with these datasets.

Low precision with Phrase Detectives and PreCo, low recall with Quiz Bowl: We observe that Phrase Detectives has a very low precision compared to all other datasets, implying that crowdworkers add more links compared to gold annotations. Our qualitative analysis reveals that PD annotators miss some valid links, splitting entities which are correctly linked together by our annotators (see Table 4). Another dataset with lower precision is PreCo, which also contains many missing links. In general, we observe more actual mistakes in PreCo and PD than in the other datasets, which is not surprising as they were not annotated by experts.[16] This result is further validated by our agreement analysis of the fiction domain (Table 5), in which `ezCoref` annotations agree far more closely with expert annotations (GUM, LitBank) than PreCo and PD. Finally, Quiz Bowl has by far the lowest recall with `ezCoref` annotations, which is expected given the difficulty with cataphora and factual knowledge (examples (c) and (e) in Table 6).

[16]That said, both PreCo and PD were additionally validated by multiple non-expert annotators.

PD	Not long after [a suitor] appeared, and as [he] appeared to be very rich and the miller could see nothing in [him] with which to find fault, he betrothed his daughter to [him] . But the girl did not care for [the man] (...). She did not feel that she could trust [him] , and she could not look at [him] nor think of [him] without an inward shudder.
PreCo	When I listened to the weather report, I was afraid to see [the advertisements] . [Those colorful advertisements] always made me crazy.

Table 4: Cases of split entities (missing links) in annotations provided with Phrase Detectives and PreCo. Instead, our crowd annotators mark all mentions as referring to the same entity in each of these examples.

Domain	Dataset	B3		
		Precision	Recall	F1
Fiction	GUM	0.982	0.921	0.950
	LitBank	0.959	0.927	0.943
	PreCo	0.805	0.963	0.877
	Phrase Detectives	0.784	0.775	0.780

Table 5: Agreement with existing datasets for fiction.

Varying the aggregation threshold τ: What is the effect of varying the aggregation threshold (τ) on precision and recall with gold annotations? Figure 5 shows that the Quiz Bowl dataset has the highest drop in recall (36% absolute drop) when increasing τ from 1 to 5.[17] This indicates that the number of unanimous clusters ($\tau = 5$) is considerably lower than the total number of clusters found individually by all annotators ($\tau = 1$); as such, our annotators heavily disagree about gold clusters in the QuizBowl dataset. We observe a similar trend in OntoNotes (26% drop in recall), whereas Phrase Detectives has the lowest drop in recall (0.07) with the increase in the number of annotators, which is expected since Phrase Detectives is crowdsourced.

5.3 What domains are most suitable for crowdsourcing coreference?

We use the B3 metric[18] (Bagga and Baldwin, 1998a) to compute IAA for each domain, excluding singletons[19] (see Table 7). We obtain the highest agreement on fiction (72.6%) and biographies (72.4%). This is because both domains contain a high frequency of pronouns (see examples *a* and

[17]We analyze variations in recall which is more interpretable than precision, since the denominator is fixed in recall when varying number of annotators.

[18]Krippendorff's alpha/kappa are other possible measures for IAA. However, prior work (Paun et al., 2022) has raised concerns over using Krippendorff's alpha/kappa for anaphora resolution. Instead, we found B3 intuitive to understand as a measure of agreement among annotators at the mention level, i.e. fraction of mentions two annotators agree should be coreferent with a given mention.

[19]IAA including singletons is much higher (Appendix A.4).

Phenomena	Dataset (Domain)	Example
Pronouns	LitBank (Fiction)	(a) A Wolf had been gorging on an animal [he] had killed, when suddenly a small bone in the meat stuck in [his] throat and [he] could not swallow [it]. [He] soon felt a terrible pain in [his] throat (...) [He] tried to induce everyone [he] met to remove the bone. "[I] would give anything, " said [he] , " if [you] would take [it] out. "
	GUM (Biographies)	(b) Despite Daniel's attempts at reconciliation, [his] father carried the grudge until [his] death. Around schooling age, [his] father, Johann, encouraged [him] to study business (...). However, Daniel refused because [he] wanted to study mathematics. [He] later gave in to [his] father's wish and studied business. [His] father then asked [him] to study in medicine.
Cataphora	QuizBowl (Quizzes)	(c) [One character in this work] is forgiven by [magenta] wife for an affair with a governess before beginning one with a ballerina. [Another character in this work] is a sickly, thin man who eventually starts dating a reformed prostitute, Marya Nikolaevna. In addition to [Stiva] and [Nikolai] , [another character in this work] (...) had earlier failed in [his] courtship of Ekaterina Shcherbatskaya.
Factual Knowledge	OntoNotes (News)	(d) The Soviet Union's jobless rate is soaring (...), [Pravda] said. Unemployment has reached 27.6 % in Azerbaijan, (...) and 16.3% in Kirgizia, [the Communist Party newspaper] said.
	QuizBowl (Quizes)	(e) (...) [another character in this work] (...) had earlier failed in [his] courtship of [Ekaterina Shcherbatskaya]. Another character in this work rejects [Ekaterina] before (...) moving to St. Petersburg. For 10 points name this work in which [Levin] marries [Kitty] , (...) a novel by Leo Tolstoy.

Table 6: Representative examples showing unique phenomena in each dataset (coreferences are color coded).

Fiction	Biographies	Opinion	Web	News	Wikipedia	Quiz
72.6	72.4	69.5	65.9	62.3	61.8	59.7

Table 7: Domain-wise IAA: B3% scores using CoNLL script (Pradhan et al., 2014), excluding singletons.

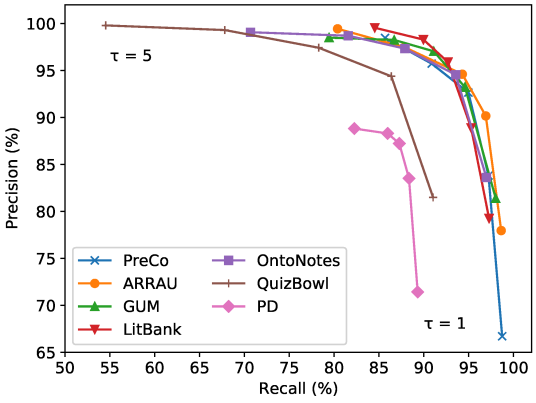

Figure 5: Agreement with gold annotations with varying voting threshold τ. $\tau = 3$ is majority voting (Figure 4). B3 scores computed with singletons included.

b in Table 6), which our annotators found easier to annotate. We also observe that the fiction domain contains many well-known children stories (e.g., Little Red Riding Hood) that are likely familiar to our annotators, which may have made them easier to annotate. Annotators have the least agreement on Quiz Bowl coreference (59.73%), as this dataset is rich in challenging cataphoras (example *c* in Table 6) and often require world knowledge about books, characters, and authors to identify coreferences (example *e* in Table 6).

5.4 Qualitative analysis

To better understand the differences in annotation quality, we conduct a manual analysis[20] of all 240 passages, comparing our `ezCoref` annotations to gold annotations from each dataset. Specifically, we look at each link that was annotated by our workers but not in the gold data, or vice versa. For each link, we determine whether crowd or the gold annotations contained a mistake, or whether the discrepancy is reasonable under specific guidelines. We find that `ezCoref` annotations contain fewer mistakes than non-expert annotated datasets (PreCo and PD), almost twice as many mistakes as those of expert datasets (OntoNotes and GUM), and seven times as many mistakes as those in the esoteric Quiz Bowl dataset (Appendix Table A2).

Disagreements and deviations from expert guidelines: As in Poesio and Artstein (2005), we identify cases of genuine ambiguity, where a mention can refer to two different antecedents. The first row of Table 8 shows an example from Dickens' *Bleak House*, where the pronoun "it" could reasonably refer to either the "fog" or the "river." Our annotators have high disagreement on this link, which is understandable given the literary analysis of Szakolczai (2016) who interprets the ambiguity of this pronoun as Dickens' way to show indeterminacy attributed to elements in the scene.[21]

We observe that generic mentions, especially generic pronouns, are almost always annotated as coreferring by crowd, while existing datasets lack consensus (Table 1). Table 8 (second row) shows an example where annotators unanimously connected all instances of generic "you." This observation is in line with Orvell et al.'s (2020) study which explains that by using the same linguistic form ("you"), one invites readers (annotators) to consider how the situation refers to them. Finally, while datasets tend to treat copulae and appositive constructions identically and annotate them

[20] By a linguist who studied guidelines of all datasets.

[21] In LitBank, the source of this passage, the pronoun "it" is annotated as referring to the "river" as only "river" is a potential markable per entity restriction (ACE entities only).

Ambiguity	[Fog] everywhere. [Fog] up [the river] , where [it] flows among green aits and meadows; [fog] down [the river] , where [it] rolls defiled among the tiers of shipping and the waterside pollutions of a great (and dirty) city. - Charles Dickens, *Bleak House*
Generic	Please , Ma'am , is this New Zealand or Australia? (and she tried to curtsey as she spoke – fancy CURTSEYING as [you] 're falling through the air! Do [you] think [you] could manage it?) - Lewis Carroll, *Alice in Wonderland*

Table 8: Examples of genuine ambiguity and generic "you" observed in our data.

in a similar way, our annotators intuitively annotate them differently. While crowdworkers almost always mark noun phrases in appositive constructions as coreferent, the noun phrases in copulae are linked by majority vote only in $\sim 35\%$ of cases.

6 Conclusion

Existing coreference datasets vary in their definition of coreferences and have been collected via complex guidelines. In this work, we investigate the quality of annotations when crowdworkers are taught only few coreference cases that are treated similarly across existing datasets. We develop a crowdsourcing-friendly coreference annotation methodology, `ezCoref` and use it to re-annotate 240 passages from seven existing English coreference datasets. We observe reasonable quality annotations were already achievable even without extensive training. On analyzing the remaining disagreements, we identify linguistic cases that crowd unanimously agree upon but lack unified treatments in existing datasets, suggesting cases the researchers should revisit when curating future unified annotation guidelines.

7 Limitations

We list some of the limitations of our study which researchers and practitioners would hopefully benefit from when interpreting our analysis. Firstly, our analysis is only applicable to the English language and how native English speakers understand coreferences. In this work, we have taken a step towards building a framework to facilitate the comparison of the crowd and expert annotations, and the variations observed in non-native speakers should be explored in future studies. Secondly, as a result of resource constraints, we limited ourselves to one set of guidelines and compared crowd annotations under these guidelines with expert annotations. Understanding the effects of various guidelines on annotator behavior is left for future research. Thirdly, even the best automatic mention detection algorithm could have errors, especially when tested out-of-domain. Despite this limitation, we decided to use an automatic method as it allows us to study annotators' behavior when a "common set of mentions" is provided. Some of the proposed solutions to address this issue are to directly crowdsource mentions or verify the automatically identified mentions via crowdsourcing (Madge et al., 2019b), which can be utilized for future collection of high-quality corpora. Finally, we also acknowledge that the tool cannot handle split-antecedents or separate tags for different relations, which we leave for future work. As a result, our approach focuses on cases of identity coreferences. However, we believe that identity coreference supported by our tool has value as an NLP tool (e.g., studying characters in narratives (Bamman et al., 2013)), allowing the collection of more in-domain annotations, necessary to advance such practical applications.

8 Ethics Statement

The data collection protocol was approved by the coauthors' institutional review board. All annotators were presented with a consent form (mentioned below) prior to the annotation. They were also informed that only satisfactory performance on the screening example will allow them to take part in the annotation task. All data collected during the tutorial and annotations (including annotators' feedback and demographics) will be released anonymized. We also ensure that the annotators receive at least $13.50 per hour. Since base compensation is per unit of work, not by time (the standard practice on Amazon Mechanical Turk), we add bonuses for workers whose speed caused them to fall below that hourly rate.

Consent Before participating in our study, we requested every annotator to provide their consent. The annotators were informed about the purpose of this research study, any risks associated with it, and the qualifications necessary to participate. The consent form also elaborated on task details describing what they will be asked to do and how long it will take. The participants were informed that they could choose as many documents as they would like to annotate (by accepting new Human Intelligence Tasks at AMT) subject to availability, and they may drop out at any time. Annotators were informed that they would be compensated in the standard manner through the Amazon Mechanical Turk crowdsourcing platform, with the amount specified in the Amazon Mechanical Turk interface. As part of this study, we also collected demographic information, including their age, gender,

native language, education level, and proficiency in the English language. We ensured our annotators that the collected personal information would remain confidential in the consent form.

Acknowledgements

We are very grateful to the crowd annotators on AMT for participating in our annotation tasks and providing positive reviews. We are grateful to Abe Handler, Aditya Jain, Anna Rogers, Julian Richardson, Kavya Jeganathan, Neha Kennard, Nishant Yadav, Timothy O'Gorman, and the UMass NLP group for several useful discussions during the course of the project. We also thank Massimo Poesio for sharing the GNOME portion of ARRAU dataset. This material is based upon work supported by National Science Foundation awards 1925548, 1814955, and 1845576, and a Google PhD Fellowship awarded to KK.

References

Amit Bagga and Breck Baldwin. 1998a. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, Volume 1*, pages 563–566.

Amit Bagga and Breck Baldwin. 1998b. 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.

David Bamman, Brendan T. O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Annual Meeting of the Association for Computational Linguistics*.

Ari Bornstein, Arie Cattan, and Ido Dagan. 2020. CoRefi: A crowd sourcing suite for coreference annotation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 205–215, Online. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2016. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2039–2046, Portorož, Slovenia. European Language Resources Association.

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.

Nicholas FitzGerald, Dan Bikel, Jan Botha, Daniel Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. MOLEMAN: Mention-only linking of entities with a mention annotation network. 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 278–285, Online. Association for Computational Linguistics.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating Zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6, New York, NY, USA. Association for Computing Machinery.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR*, abs/1803.09010.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. 2014. CROMER: A tool for cross-document event and entity coreference. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3204–3208, Reykjavik, Iceland. European Language Resources Association.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the Training Wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1108–1118, Denver, Colorado. Association for Computational Linguistics.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Play the Language: Play Coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 209–212, Suntec, Singapore. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Jeff Howe. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business*. London, England: Random House Books.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, and Massimo Poesio. 2019. Wormingo: A 'true gamification' approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7.

Osman Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, Silviu Paun, and Massimo Poesio. 2020. Aggregation driven progression system for GWAPs. In *Workshop on Games and Natural Language Processing*, pages 79–84, Marseille, France. European Language Resources Association.

Mateusz Kopeć. 2014. MMAX2 for coreference annotation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 93–96, Gothenburg, Sweden. Association for Computational Linguistics.

Frédéric Landragin, Thierry Poibeau, and Bernard Victorri. 2012. ANALEC: A new tool for the dynamic annotation of textual data. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 357–362, Istanbul, Turkey. European Language Resources Association.

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. 2010. Annotating large email datasets for Named Entity Recognition with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79, Los Angeles. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8320–8331, Online. Association for Computational Linguistics.

Robert L Logan IV, Andrew McCallum, Sameer Singh, and Dan Bikel. 2021. Benchmarking scalable methods for streaming cross document entity coreference. 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 4717–4731, Online. Association for Computational Linguistics.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019a. Making text annotation fun with a clicker game. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, New York, NY, USA. Association for Computing Machinery.

Chris Madge, Juntao Yu, Jon Chamberlain, Udo Kruschwitz, Silviu Paun, and Massimo Poesio. 2019b. Progression in a Language Annotation Game with a Purpose. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 77–85.

Josh Aaron Miller, Uttkarsh Narayan, Matthew Hantsbarger, Seth Cooper, and Magy Seif El-Nasr. 2019. Expertise and engagement: Re-designing citizen science games with players' minds in mind. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–11.

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.

Christoph Mueller and Michael Strube. 2001. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Bruno Oberle. 2018. SACR: A drag-and-drop based tool for coreference annotation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Ariana Orvell, Ethan Kross, and Susan A. Gelman. 2020. "You" speaks to me: Effects of generic-you in creating resonance between people and ideas. *Proceedings of the National Academy of Sciences*, 117(49):31038–31045.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*, volume 1 of *Synthesis Lectures on Human Language Technologies*. Springer, Cham.

Massimo Poesio and Ron Artstein. 2005. Annotating (Anaphoric) Ambiguity. In *Proceedings of the corpus linguistics conference*.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1).

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1778–1789. Association for Computational Linguistics.

Massimo Poesio, Amir Zeldes, Anna Nedoluzhko, Sopan Khosla, Ramesh Manuvinakurike, Nafise Moosavi, Vincent Ng, Maciej Ogrodniczuk, Sameer Pradhan, Carolyn Rose, Michael Strube, Juntao Yu, Yulia Grishina, Yufang Hou, and Fred Landragin. 2021. Universal anaphora 1.0. `https://sites.google.com/view/universalanaphora/`. Accessed: 2021-10-30.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. 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.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. 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*, pages 101–108, Online. Association for Computational Linguistics.

James Ravenscroft, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. 2021. CDˆ2CR: Coreference resolution across documents and domains. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 270–280, Online. Association for Computational Linguistics.

Nils Reiter. 2018. CorefAnnotator - A new annotation tool for entity references. In *Abstracts of EADH: Data in the Digital Humanities*.

Rion Snow, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Árpád Szakolczai. 2016. *Novels and the Sociology of the Contemporary*. Routledge, Milton Park, Abingdon, Oxon New York, NY.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for Word Sense Labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany. Association for Computational Linguistics.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel-aati Hawwary, Marcus Mitchell, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2012. Ontonotes Release 5.0. `https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf`. Accessed: 2022-01-15.

Antoine Widlöcher and Yann Mathet. 2012. The Glozz platform: A corpus annotation and mining tool. In *2012 ACM symposium on Document Engineering*.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the*

51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Carretero Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2022. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and Wikipedia texts. *Computing Research Repository*, arXiv:2210.05581.

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

A Appendix

A.1 Details of our crowdsourced data

Table A1 mentions all datasets that we re-annotate in this work with their breakdown based on domains, number of documents, passages, tokens and mentions annotated.

Dataset	Domain	#Docs	#Passages	#Tokens	#Mentions
OntoNotes	News	6	30	4923	1365
	Weblogs	5	20	3452	1001
	Opinion	12	20	3861	1157
LitBank	Fiction	4	30	5455	1494
QuizBowl	Quizzes	20	20	3304	1083
ARRAU	News	3	20	3336	885
GUM	Biographies	4	20	3422	1119
	Fiction	4	20	3299	1008
Phrase	Wikipedia	7	20	3509	1003
Detectives	Fiction	4	20	4007	1063
PreCo	Opinion	7	9	1692	495
	News	4	8	1318	369
	Fiction	2	2	378	105
	Biographies	1	1	152	53
Total	All	83	240	42108	12200

Table A1: All datasets analyzed in this work with their breakdown based on domains, number of documents, passages, tokens and mentions annotated.

A.2 Manual Qualitative Analysis

Dataset	Mistakes (our)	Mistakes (gold)
PD (silver)	22	76
OntoNotes	81	49
PreCo	12	33
GUM	48	25
ARRAU	33	16
LitBank	21	13
QuizBowl	67	10

Table A2: Number of mistakes in our crowd annotations vs. gold datasets, obtained through a manual analysis.

A.3 Detailed Mention Detection Algorithm

- We identify all noun phrases using the Stanza dependency parser (Qi et al., 2020). For each word with a noun-related part-of-speech tag,[22] we recursively traverse all of its children in the dependency graph until a dependency relation is found in a `whitelist`.[23] The maximal span considered as a candidate mention thus covers all words related by relations in the `whitelist`.

- Possessive nominal modifiers are also considered as candidate mentions. For instance, in the sentence "Mary's book is on the table," we consider both "Mary" and "Mary's book" as mentions.

- Modifiers that are proper nouns in a multi-word expression are considered as mentions. For instance, in "U.S. foreign policy," the modifier "U.S." is also considered as a mention.

- All conjuncts, including the headword and other words depending on it via the conjunct relation, are considered mentions in a coordinated noun phrase. For instance, in the sentence, "John, Bob, and Mary went to the party.", the detected mentions are "John," "Bob," "Mary," and the coordinated noun phrase "John, Bob, and Mary."

- Finally, we remove mentions if a larger mention with the same headword exists. We allow nested spans (e.g., [[my] hands]) but merge any intersecting spans into one large span (e.g, [western [Canadian] province] is merged into [western Canadian province]).

[22] Pronouns, nouns, proper nouns, and numbers.

[23] The `whitelist` includes all multi-word expression relations (i.e., compound, flat, and fixed) and modifier relations (i.e., determiners, adjectival modifiers, numeric modifiers, nominal modifiers, and possessive nominal modifiers).

A.4 Inter-Annotator Agreement Among Our Annotators Across Domains

Figure 6 illustrates agreement among our annotators computed with B3 scores including singletons.

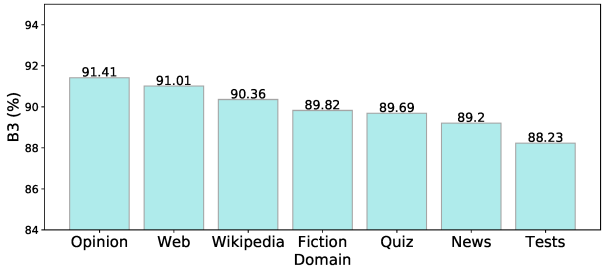

Figure 6: Inter Annotator Agreement across different domains. B3 scores with Singletons included.

A.5 Another illustrative example

An example of a single sentence annotated by two datasets, OntoNotes and ARRAU. These annotations differ widely from each other in kinds of mentions and links between mentions.

OntoNotes: [Lloyd's, once a pillar of [the world insurance market]e1,]e2 is being shaken to [its]e2 very foundation.

ARRAU: [Lloyd's, once [a pillar of [the world [insurance]e3 market]e2]eS1]e1, is being shaken to [[its]e1 very foundation]eS2.

System	Annotate all clusters	Pre-identified Mentions	Open Source	Webapp	Coref only	Keyboard and Mouse	MTurk Tested	Non-expert Terminology	Nested Span Support	Interactive Tutorial
Stenetorp et al. (2012)	✓	✗	✓	✓	✗	✗	✗	✓	✗*	✗
Widlöcher and Mathet (2012)	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗
Landragin et al. (2012)	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗
Yimam et al. (2013)	✓	✗	✓	✓	✗	✗	✗*	✗	✓	✗
Poesio et al. (2013)	✗	✓	✗	✓	✓	✗	✗	✓	✓	✓
Girardi et al. (2014)	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗
Mueller and Strube (2001)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
Kopeć (2014)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
Guha et al. (2015)	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗
Oberle (2018)	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗
Reiter (2018)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
Bornstein et al. (2020)	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Prodigy*	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓
`ezCoref` (this work)	✓	✓	✓*	✓	✓	✓	✓	✓	✓	✓

Table A3: A comparison of different coreference annotation tools. (* — `ezCoref` code will be open-sourced upon paper publication; Stenetorp et al. (2012) did not implement nested spans originally, but later added them with limited functionality. Yimam et al. (2013) have APIs for CrowdFlower integration, but suggest expert annotators.).
*Accessible at: `https://prodi.gy/`

Tutorial feedback from our crowd annotators
1. This was a really interesting task. The tutorial was very clear and easy to understand. I think it was very helpful when I completed the final passage.
2. Very great tutorial, I loved how it walked me through each and every step making sure I understood.
3. excellent interface and very precise instructions! out of curiousity, what is the time-frame and scale for this project? several weeks? months? hundreds or thousands of hits? I have a ton of projects during the autumn normally but will definitely make time for this if it's going to be around for more than a day or two. Looking forward to working with you folks if possible!
4. I actually enjoyed this. Thank you for the opportunity.
5. it was interesting a bit difficult but overall gave a lot of feedback necessary to do a good job.
6. I loved the tutorial and the layout. I am still a little bit unsure about a couple of the entities and hope I got it right. For example: would 'legs' be in 'his' because it refers to that person? I wasn't sure and made them separate.
7. I loved how this tutorial was set up. It was easy to use and made me very interested in doing the actual HITs. It would have been nice to be able to print out a quick reference guide or something, so we could refer to the instructions from before while we completed the final task. I don't think it would be needed for very long after starting the real HITs, but it would still be nice to have.
8. On the last test section, there was no place for feedback. There was a section that said ""it was getting dark"" ""It was getting late"" Both of those refer to a time of day, but one is light, one is the hour, so I marked them as different. Not sure of how broad or narrow we need to be when justifying ""same"" entities, as there is an argument either way.
9. I just wanted to say that I really appreciated how efficiently put together and clear this tutorial was.
10. This was a unique task. Thank you.
11. I feel much better with the help and feedback. It was interesting and definitely way different in a good way than the usual survey. I did my best and I hope I did well enough. Keep safe and Happy Holidays no matter what happens.

Table A4: Some of the comments received from our annotators after completing the tutorial. We received overwhelmingly positive feedback; annotators sometimes also mentioned cases they found confusing.

Coreference Tutorial

Welcome!

This is a **paid tutorial** for the **"Large-Scale Coreference Annotation Task."**

In this tutorial you will learn how to annotate **coreferences**, that is, words and phrases that refer to the same people or things.

Upon completing the tutorial, you will get **a completion code.** You **MUST enter this code** in the textbox below and **submit the HIT** in order to receive the payment.

Depending on your performance, you might be invited to participate in our "Large-Scale Coreference Annotation Task."

Before proceeding to the tutorial, please **fill in the following survey**:

What is your **gender**?

_____

What is your **age**?

_____

What is your **native language**?

_____

How is your **English level**?

- Beginner
- Intermediate
- Advanced (near native)
- Native speaker

What is your **education level**?

- Primary
- Secondary
- College
- Graduate School

Click this link to begin.

[OPTIONAL] We would love to hear **your feedback** about **this tutorial.**

Submit your code below:

Submit

Figure 7: Screenshot of tutorial task invitation on AMT with detailed instructions.

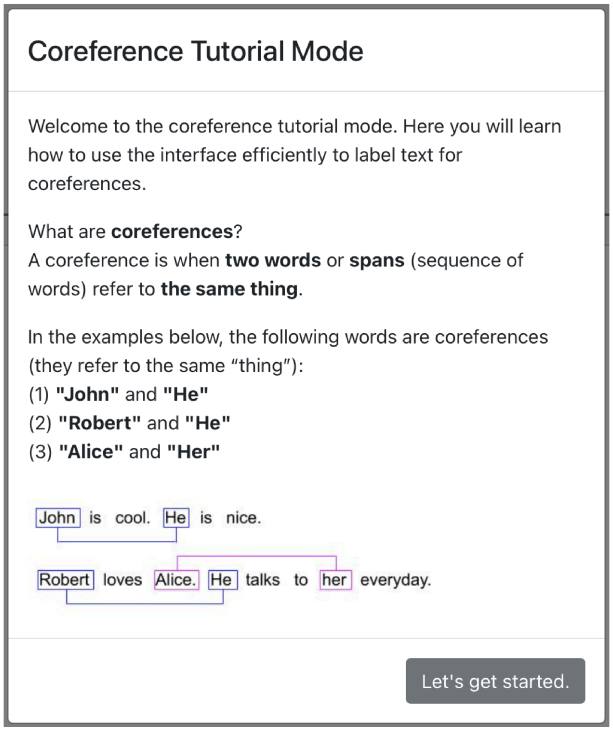

Figure 8: Tutorial Interface (Introductory prompt)

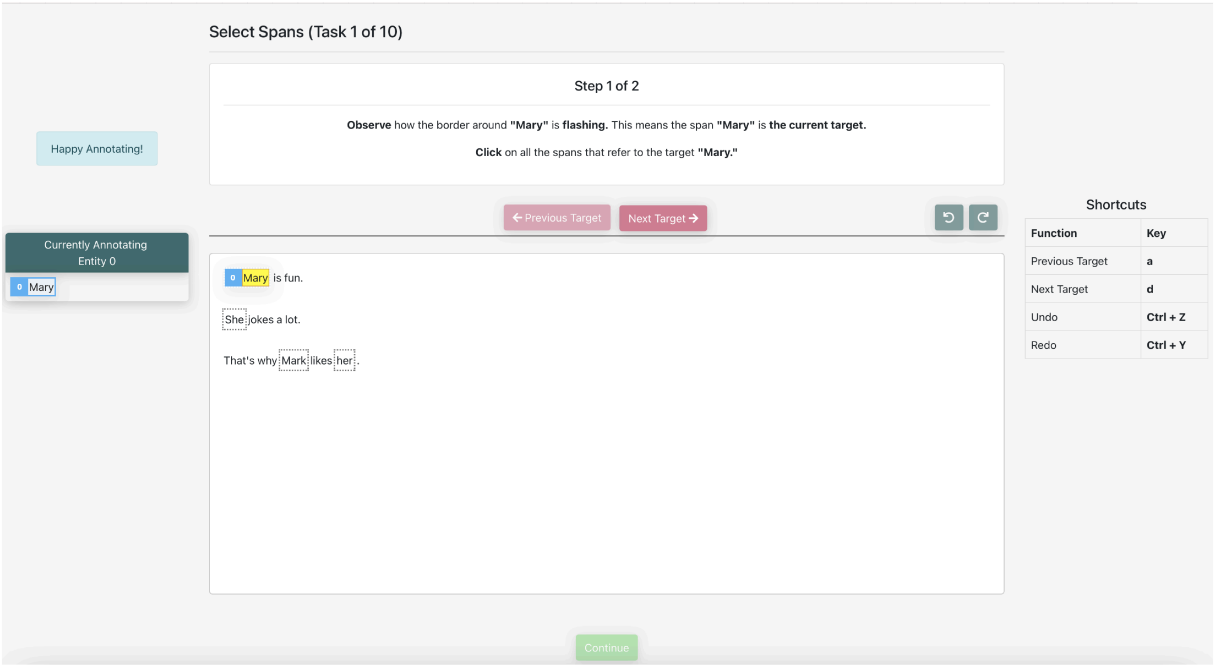

Figure 9: Tutorial interface: A sample prompt teaching tool functionality.

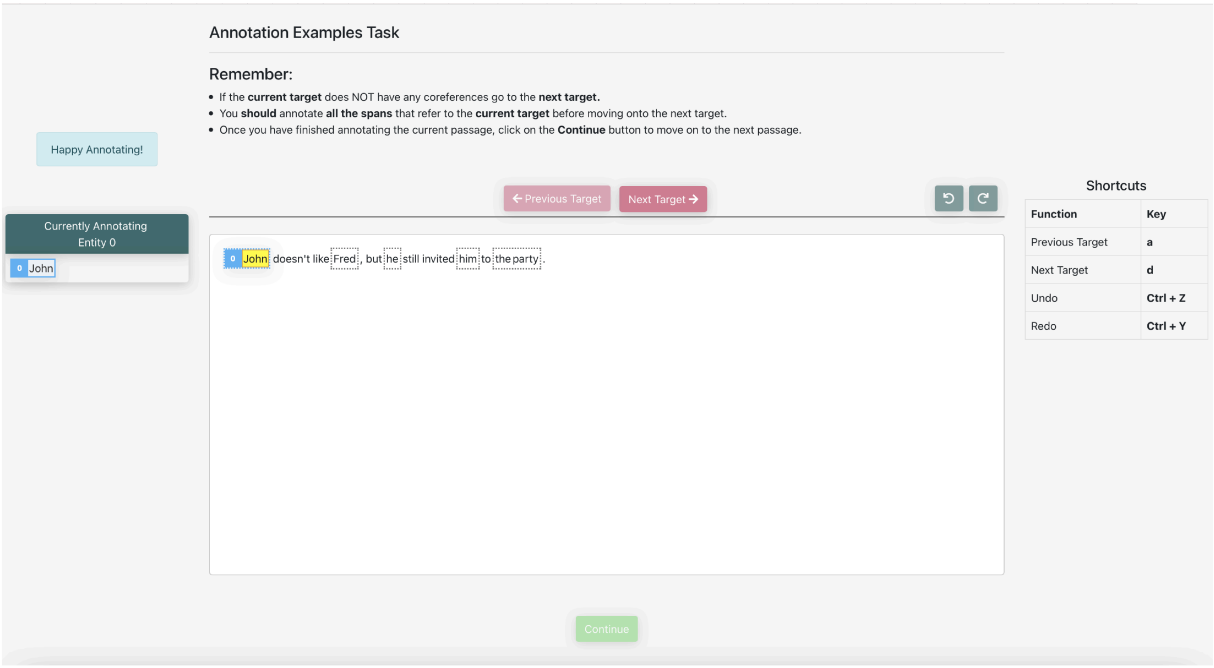

Figure 10: Tutorial interface: A sample prompt teaching basic coreferences.

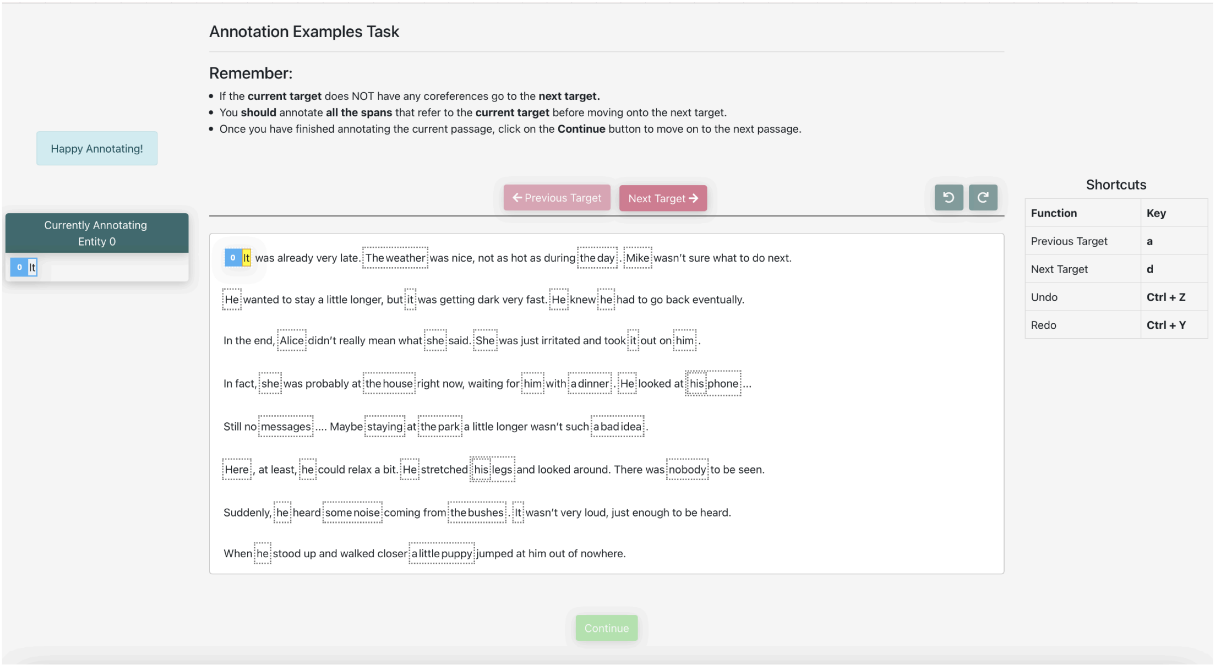

Figure 11: Tutorial interface: quality control example.

Coreference Annotation Task

Welcome to the coreference annotation task. In this task you will be asked to annotate a short paragraph for coreferences. If you need to review the tutorial, please follow this link.

What are **coreferences**?

A coreference is when **two words** or **spans** (sequence of words) refer to **the same thing**.

In the examples below, the following words are coreferences (they refer to the same "thing"):

(1) "**John**" and "**He**"

(2) "**Robert**" and "**He**"

(3) "**Alice**" and "**Her**"

John is cool. He is nice.

Robert loves Alice. He talks to her everyday.

Click this link to begin annotation.

[OPTIONAL] We would love to hear **your feedback**. Let us know if anything was unclear or particularly challenging.

Submit your code below:

Submit

Figure 12: Annotation task invite on AMT with detailed instructions