`ezCoref`: A Scalable Approach for Collecting Crowdsourced Annotations for Coreference Resolution

Anonymous ACL submission

Abstract

Large-scale high-quality corpora are critical for advancing research in coreference resolution. Coreference annotation is typically time-consuming and expensive, since researchers generally hire expert annotators and train them with an extensive set of guidelines. Crowdsourcing is a promising alternative, but coreference includes complex semantic phenomena difficult to explain to untrained crowdworkers, and the clustering structure is difficult to manipulate in a user interface. To address these challenges, we develop and release `ezCoref`, an easy-to-use coreference annotation tool and annotation methodology that facilitates crowdsourced data collection across multiple domains, currently in English. Instead of teaching crowdworkers how to handle non-trivial cases (e.g., near-identity coreferences), `ezCoref` provides only a minimal set of guidelines sufficient for understanding the basics of the task. To validate this decision, we deploy `ezCoref` on Mechanical Turk to re-annotate 240 passages from seven existing English coreference datasets across seven domains, achieving an average rate of 2530 tokens per hour, for one annotator. This paper is the first to compare the quality of crowdsourced coreference annotations against those of experts, and to identify where their behavior differs to facilitate future annotation efforts. We show that it is possible to collect coreference annotations of a reasonable quality in a fraction of time it would traditionally require.

1 Introduction

Coreference resolution is the task of identifying all textual expressions that refer to the same discourse entity in a given document, and thus grouping such coreferent expressions (mentions) into clusters (entities). Systems trained to solve this task are often an integral component of the preprocessing pipeline for many downstream tasks, such as summarization (Azzam et al., 1999; Steinberger et al., 2007), question answering (Vicedo and Fernández, 2000; Dhingra et al., 2018), and machine translation (Hardmeier, 2012; Bawden et al., 2018). Modern coreference systems are implemented via data-hungry neural network models (e.g., Lee et al., 2017; Moosavi and Strube, 2018; Joshi et al., 2019) trained on large-scale expert-annotated datasets such as OntoNotes (Weischedel et al., 2013).

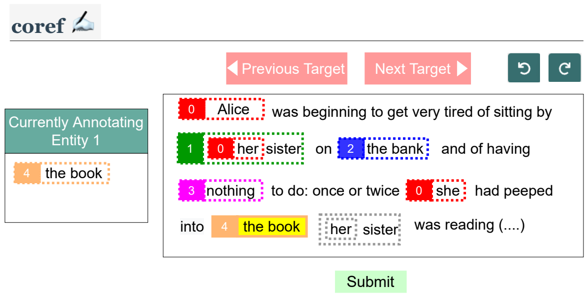


Figure 1: Part of the `ezCoref` interface (§3)

Acquiring these datasets has traditionally been difficult, expensive and time-consuming, requiring linguists trained in fine-grained annotation schemas (e.g., Hovy et al., 2006; Poesio and Artstein, 2008; Uryupina et al., 2019). As such, coreference datasets exist only for a small set of languages (mostly English) and even then only for limited domains (mainly news and fiction). Furthermore, these datasets differ widely in their annotation guidelines, resulting in inconsistent annotations across languages and domains, with challenging cross-lingual and cross-domain issues (Poesio et al., 2021).

Is it possible to use crowdsourcing and non-expert annotators to generate high-quality coreference data? As in other types of linguistic annotations, crowdsourcing could reduce costs (Snow et al., 2008), allowing for larger scale datasets. Furthermore, by using a standard platform like Amazon Mechanical Turk, this approach is accessible to a wider range of researchers, who may wish to collect new data for different corpora or even languages.

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
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
GUM (**Zeldes, 2017**)	Multiple (25, 6K, 20K)	Experts (Linguistics Students)	Manual	Yes	None	Coref (Sub-Types)	Coref (Sub-Type)	Yes	None
PreCo (**Chen et al., 2018a**)	Multiple*** (38K, 3.58M, 12.5M)	Non-Expert, Non-Native	Manual**	Yes	None	Coref	Coref	Yes	None
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
QuizBowl (**Guha et al., 2015**)	Single (400, 9.4K, 50K)	Domain Experts	Manual & CRF*	Yes	Characters, Books, Authors*	Coref	Coref	If Applicable	None
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., 2018b). ***See Footnote 8.

To this end, we develop `ezCoref`, a crowdsourced coreference annotation platform that is intuitive and easy to use for crowdworkers, and open-source for other researchers to utilize.[1] Unlike existing crowdsourced coreference efforts (Chamberlain et al., 2016; Bornstein et al., 2020; Li et al., 2020), `ezCoref` simplifies the annotation task for workers by using automatically detected mention boundaries[2], is easily integrable with platforms like Amazon Mechanical Turk, and offers a short, effective, crowd-oriented interactive tutorial[3].

With this new interface, we turn to the question of how to define coreference to untrained crowdworkers. Expertly-collected datasets such as OntoNotes explain what should and should not be considered coreference via a lengthy set of guidelines (Weischedel et al., 2012) that covers many complex linguistic details (e.g., how to deal with head-sharing noun phrases, which premodifiers can and cannot corefer, or how to annotate generic mentions). Even if it was feasible to teach such guidelines to crowdworkers, existing coreference datasets differ widely in terms of what is considered as a mention and what types of links should be annotated (Poesio et al. (2021) and Table 1), so it is unclear what standards ought to be used.

Instead, we explore whether we can collect high-quality coreference data with a *minimal* set of guidelines, only illustrating basic phenomena like pronoun resolution (Table 3). We use `ezCoref` to re-annotate a subset of documents from seven different English coreference datasets. Our crowdworkers obtain high agreement with most of the expert annotations, and that the quality of our annotations is higher than that of previous crowdsourced efforts. We conclude with a qualitative analysis of our annotators' behavior and the types of annotation decisions they make, which we hope will inform future research into coreference platform development and guideline construction.

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.[4] 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). Coreference datasets annotated by non-experts include those created by part-time non-native English speakers (PreCo; Chen et al. (2018a)), domain but not linguistic experts (QuizBowl; Guha et al. (2015)), and gamified crowdsourcing without

[1]Our platform's code and collected data is available in supplementary materials, and will be released publicly after blind review.

[2]The syntax-based mention detector is our system's only English-specific component.

[3]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.*"

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

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)	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗
Kopeć (2014)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
Guha et al. (2015)	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗
Oberle (2018)	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗
Reiter (2018)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗
Bornstein et al. (2020)	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
ezCoref (this work)	✓	✓	✓*	✓	✓	✓	✓	✓	✓	✓

Table 2: 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.)

financial compensation (Phrase Detectives; Chamberlain et al. (2016)).

Coreference annotation tools: Several coreference annotation tools with similar features to ezCoref have already been developed (Table 2). 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; Reiter, 2018), or have complicated 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; Oberle, 2018). Closest to our work 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.

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 Amazon Mechanical Turk 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.

User interface: We spent two years iteratively designing, implementing, and user testing the interface in order to make it as simple and crowdsourcing-friendly as possible (Figure 1). 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, and undo and redo actions were added to allow error correction. Finally, ezCoref provides a side panel to show mentions of the entity currently being annotated, which helps 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 familiarizes them with the mechanics of the annotation tool, then trains them in a minimal set of annotation guidelines (Table 3). The tutorial concludes with a quality control example to exclude poor quality

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 3: Phenomena explained in our tutorial.

annotators.[5] Training examples, feedback, and annotation guidelines can be easily customized using a simple JSON configuration schema.

Annotation workflow: The annotators are presented with one passage (or "document") at a time (Figure 1), 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.

In this work, we divide an initial document into a sequence of shorter passages of complete sentences, on average 175 tokens, since 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: To simplify the annotation task for crowdworkers, we decide to automatically annotate mentions instead of forcing workers to mark mention boundaries in the text. This approach considerably reduces the annotator's effort while speeding annotation; however, it relies heavily on the performance of the mention detection algorithm, which also can detect non-referring expressions (e.g., "hand" in "on the other hand"). Table 1 shows that existing datasets use many methods and criteria to identify mentions, from completely manual to semi-automatic and fully automatic procedures. Instead of choosing an existing standard, we implement a simple automatic mention detection algorithm that yields a high average recall over all seven of these datasets. We consider all noun phrases (including proper nouns, common nouns, and pronouns) as markables, extracting them using the Stanza[7] dependency parser (Qi et al., 2020). We allow for nested mentions and proper noun premodifiers (e.g., [U.S.] in "U.S. policy"). We also include all conjuncts with the entire coordinated noun phrase ([Mark], [Mary], as well as [Mark and Mary], are all considered mentions); see Appendix A.2 for more details.

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

To study annotator behavior in our setup, we deploy `ezCoref` on the Amazon Mechanical Turk (AMT) crowdsourcing platform to re-annotate 240 passages from seven existing datasets, covering seven unique domains. We compare our workers' annotations to each other across domains, compare them to the previous gold standard annotations, and conduct our own qualitative analysis as well. In total, we collect annotations for 12,200 mentions and 42,108 tokens.

4.1 Experimental Setup

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., 2018a), 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 se-

[5]Examples of the tutorial interface and the quality control example are provided in Appendix A.6.

[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]version 1.3.0

[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 (Appendix A); we use our manual categories for domain analysis.

lect domain(s) to re-annotate so that we cover a broad range of domains. 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 annotate 240 passages, collecting five annotations per passage to measure inter-annotator agreement.

Procedure: We first launch an annotation tutorial (paid $4.50) and recruit the annotators on the AMT platform.[9] As the goal of our study is to understand what crowd annotators perceive as coreference and to identify instances of genuine ambiguity, we train our annotators, providing them with minimal guidelines. We carefully draft our training examples to include only cases which are considered as coreference across all the existing datasets (i.e., we exclude copular expression, appositives, etc.). The objective is to teach crowd annotators the broad definition of coreference while leaving space for different interpretations of cases which are ambiguous or resolved differently across the existing datasets. 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.

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. The main annotation task was paid $1 per passage, resulting in the total cost of $1440.[11]

[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.

5 Analysis

In this section, we perform a quantitative and qualitative analysis of our crowdsourced coreference annotations. First, we evaluate the performance of our mention detection algorithm, comparing it to gold mentions across seven datasets. Next, we measure the quality of our annotations (via inter-annotator agreement between our crowdworkers) 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 mentions‘ boundaries. Hence, the boundaries for the same mention may differ. To fairly compare our mentions with the gold standards, we employ a headword-based comparison. In particular, 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. We consider two mentions as the same if their respective headwords match.

Table 4 compares our mention detector to the gold mentions in existing datasets. Our method obtains high recall across most datasets (>0.90). 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] Comparing precision, 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). As a result, our algorithm identifies more mentions than in the original datasets, which also allows us to discover new entities. For the remaining datasets, the precision is >0.80, suggesting that the algorithm identifies most of the relevant mentions. Finally, we compare mention density (number of mentions per token) between our detector and existing datasets, and find that while gold mention density varies considerably across the seven datasets due to their differing mention criteria (Table 1), it

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

[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 4: 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.

remains roughly consistent across all datasets when using our method.

5.2 What domains are most suitable for crowdsourcing coreference?

Which domains yield the highest inter-annotator agreement (IAA) between our crowdworkers?

We use the B3 metric[13] (Bagga and Baldwin, 1998a) to compute IAA for each domain, excluding singletons[14] (see Table 6). 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 *b* in Table 5), 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 5) and often require world knowledge about books, characters, and authors to identify coreferences (example *e* in Table 5).

5.3 Agreement with Existing Datasets

Having established relatively good agreement amongst our workers on most domains, we now turn to a different question: how well do crowdsourced annotations from `ezCoref` agree with gold 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. 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.[15] After aggregating our `ezCoref` annotations, we compare these annotations with gold annotations across the seven datasets using B3 scores (precision, recall, and F1), as illustrated in Figure 2.

High agreement with OntoNotes, GUM, LitBank, ARRAU: Our annotators achieve the highest precision with OntoNotes, 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 `ezCoref` facilitates high quality annotations from untrained crowdworkers.

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 7). 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 8), in which `ezCoref` annotations agree far more closely with

[13] We also computed Krippendorff's α for inter-annotator agreement and obtained similar results.

[14] The agreement including singletons is substantially higher. The exact numbers are provided in Appendix A.3.

[15] 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.

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

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 (Quizes)	(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 5: Examples taken from respective datasets to illustrate their unique phenomena. Coreferent mentions are marked with same color in each example.

Fiction	Bio	Opinion	Web	News	Wiki	Quiz
72.6	72.4	69.5	65.9	62.3	61.8	59.7

Table 6: Inter Annotator Agreement (B3 %) across different domains. B3 scores are computed in accordance with the CoNLL script (Pradhan et al., 2014), excluding singletons. Bio (Biographies); Wiki (Wikipedia).

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

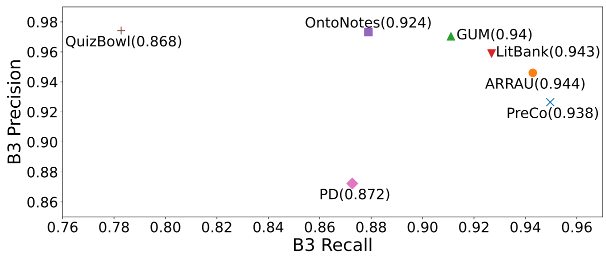


Figure 2: Agreement with gold annotations across datasets. B3 (F1) scores shown in parentheses are computed with singletons included.

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 8: Agreement with existing datasets for fiction domain.

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.

Varying the aggregation threshold τ: What is the effect of varying the aggregation threshold (τ) on precision and recall with gold annotations? Figure 3 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.

[17]We analyze variations in recall since it is more interpretable than precision, given that the denominator is fixed in recall with a variable number of annotators.

5.4 Qualitative analysis

To better understand the differences in annotation quality, we conduct a manual analysis of all 240 passages in our experiment, 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 our workers or the gold annotations contained a mistake, or whether the discrepancy is reasonable under specific guidelines. Table 9 shows that our `ezCoref` annotations contain

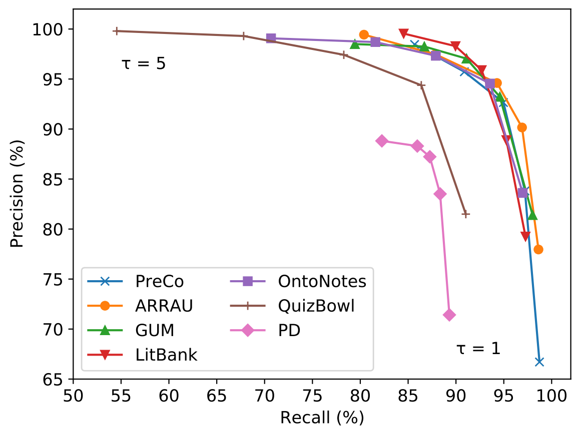


Figure 3: Agreement with gold annotations when varying voting threshold τ. $\tau = 3$ is majority voting (Figure 2). B3 scores are computed with singletons included.

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

Table 9: Number of mistakes in our crowd annotations vs. gold datasets, obtained through a manual analysis.

fewer mistakes than non-expert annotated datasets such as PreCo and PD, but there are almost twice as many mistakes as those of expert datasets such as OntoNotes and GUM, and seven times as many mistakes as those in the esoteric Quiz Bowl dataset.

Qualitative examples of disagreement: As in Poesio and Artstein (2005), we identify some cases of genuine ambiguity, where a mention can refer to two different antecedents. The first row of Table 10 shows one such 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) interprets the ambiguity of this pronoun as Dickens' way to show indeterminacy attributed to elements in the scene.[18] We also observe that generic mentions, especially generic pronouns, are almost always annotated as coreferring by our annotators. The second row of Table 10 shows one example of such a case, where annotators unanimously connected all instances of generic "you." While generic pronouns are usually regarded as non-referring (Huddleston, 2002), they retain something of their specific quality as personal pronouns (Quirk, 1985). Finally, while datasets tend to treat copulae and appositive constructions identically and annotate them in a similar way, our annotators intuitively annotate them differently. Although they almost always mark noun phrases in appositive constructions as coreferents, the noun phrases in copulae are linked by majority vote only in about 35% of the cases.

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

6 Conclusion

In this work, we present `ezCoref`, an intuitive and easy-to-use annotation tool to collect crowdsourced annotations for coreference resolution. Using `ezCoref`, we re-annotate a subset of documents from seven different English coreference datasets, each of which was created using a different set of complex linguistic guidelines. In contrast, our `ezCoref` re-annotation aims to collect collect high-quality coreference data with a *minimal* set of guidelines. Our results show that crowdworkers obtain high agreement with many expert annotations (e.g., GUM, ARRAU) and that our annotation quality is better than previous crowdsourced efforts (e.g., Phrase Detectives). We hope our `ezCoref` tool and observations will inform future research into coreference platform development and guideline construction.

7 Ethics Statement

The data collection protocol was approved by the coauthors' institutional review board. All annotators were presented with a consent form 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 demograph-

[18]In LitBank, where this passage came from, the pronoun "it" is annotated as referring to the "river" as only "river" is a potential markable per entity restriction (selected ACE entities only).

ics) 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.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

Amit Bagga and Breck Baldwin. 1998a. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

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 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. 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.

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 (LREC'16)*, pages 2039–2046, Portorož, Slovenia. European Language Resources Association (ELRA).

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018a. 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.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018b. 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.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48, New Orleans, Louisiana. 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.

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 (LREC'14)*, pages 3204–3208, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Christian Hardmeier. 2012. Discourse in statistical machine translation : A survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique.A journal of linguistics, psycholinguistics and computational linguistics*, 11.

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.

Rodney Huddleston. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK New York.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

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 (LREC'12)*, pages 357–362, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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.

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 (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Massimo Poesio and Ron Artstein. 2005. Annotating (anaphoric) ambiguity.

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

Massimo Poesio, 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, 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*.

Randolph Quirk. 1985. *A Comprehensive grammar of the English language*. Longman, London New York.

James Ravenscroft, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. 2021. CD^2CR: Co-reference 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, Daniel Jurafsky, and Andrew 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, Honolulu, Hawaii. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two uses of anaphora resolution in summarization. *Inf. Process. Manage.*, 43(6):1663–1680.

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.

Arpad 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.

José L. Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 555–562, Hong Kong. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

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 *DocEng '12*.

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.

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

A Appendix

A.1 Details of our crowdsourced data

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 Detectives	Wikipedia	7	20	3509	1003
	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

A.2 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,[19] we recursively traverse all of its children in the dependency graph until a dependency relation is found in a `whitelist`.[20] The maximal span considered as a candidate mention thus covers all words related by relations in the `whitelist`.

[19] pronouns, nouns, proper nouns, and numbers.

[20] 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).

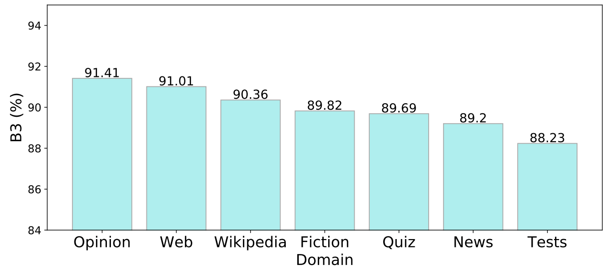


Figure 4: Inter Annotator Agreement across different domains. B3 scores with Singletons included.

- 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]).

A.3 Inter-Annotator Agreement Among Our Annotators Across Domains

Figure 4 illustrates agreement among our annotators computed with B3 scores including singletons.

A.4 An 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.

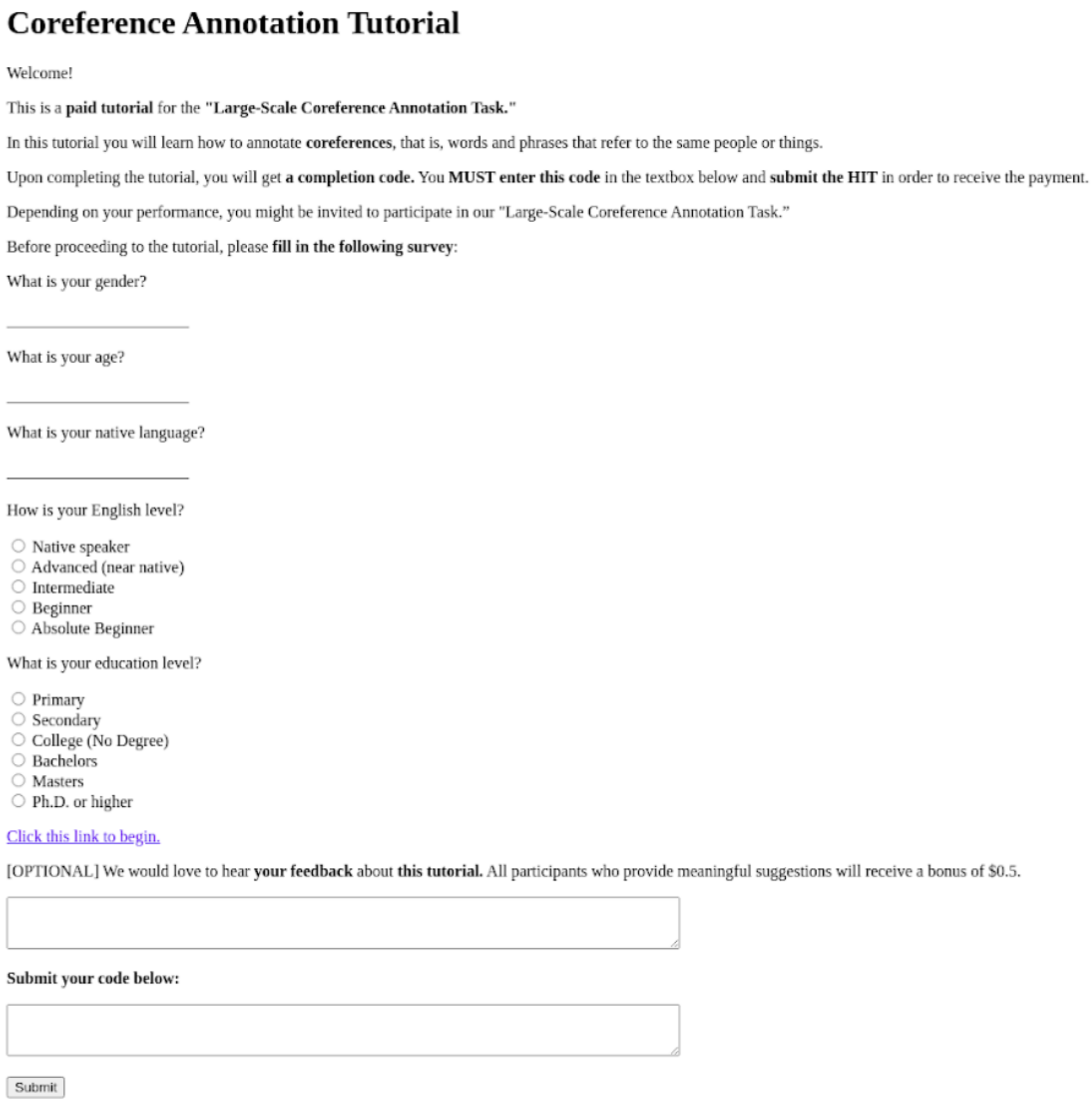
OntoNotes: [Lloyd's, once a pillar of [the world insurance market]e1,]e2 is being shaken to [its]e2 very foundation.

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

A.5 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.

A.6 Details of Tutorial



Coreference Annotation 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?

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

What is your education level?

- Primary
- Secondary
- College (No Degree)
- Bachelors
- Masters
- Ph.D. or higher

Click this link to begin.

[OPTIONAL] We would love to hear **your feedback** about **this tutorial.** All participants who provide meaningful suggestions will receive a bonus of $0.5.

Submit your code below:

Submit

Figure 5: Screenshot of tutorial task invitation on AMT with detailed instructions.

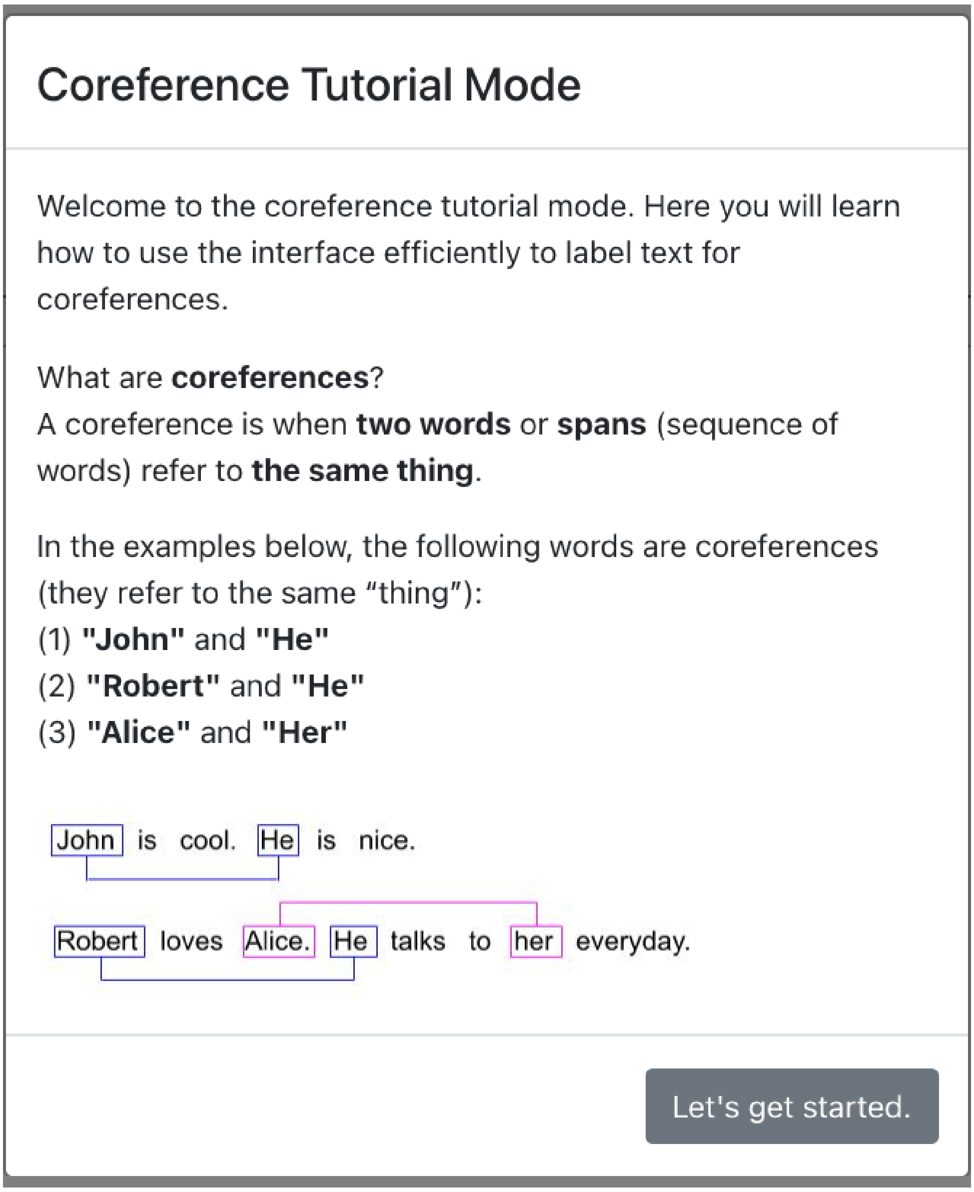


Figure 6: Tutorial Interface (Introductory prompt)

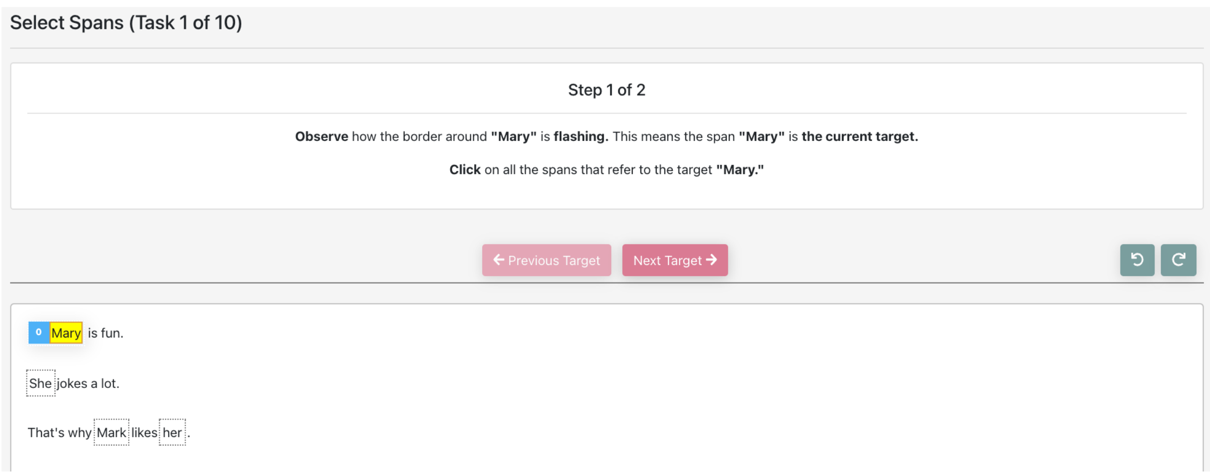


Figure 7: Tutorial interface: A sample prompt teaching tool functionality.

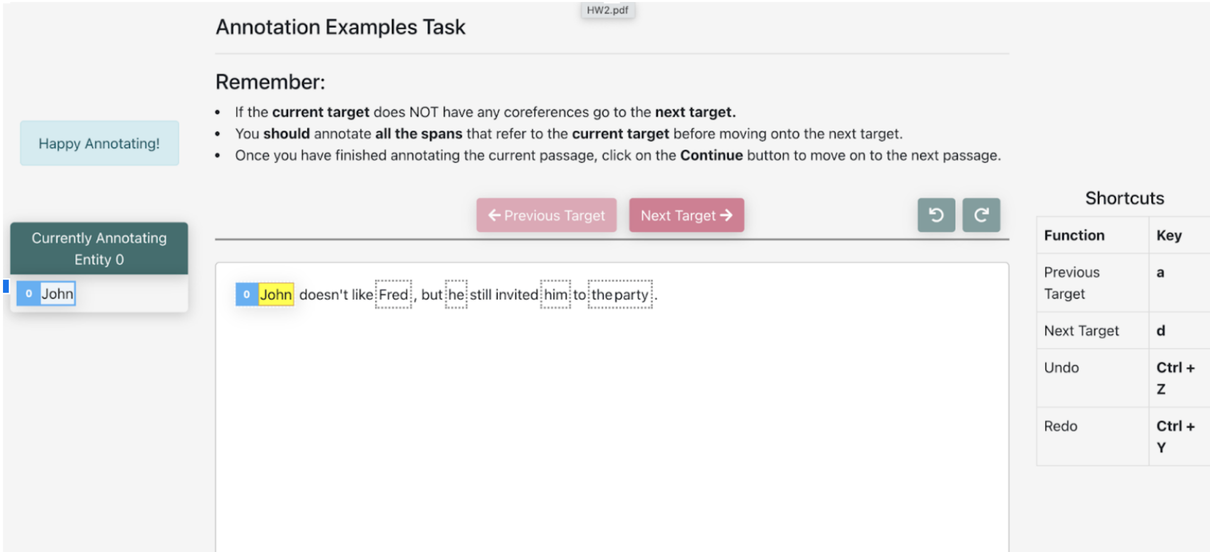


Figure 8: Tutorial interface: A sample prompt teaching basic coreferences.

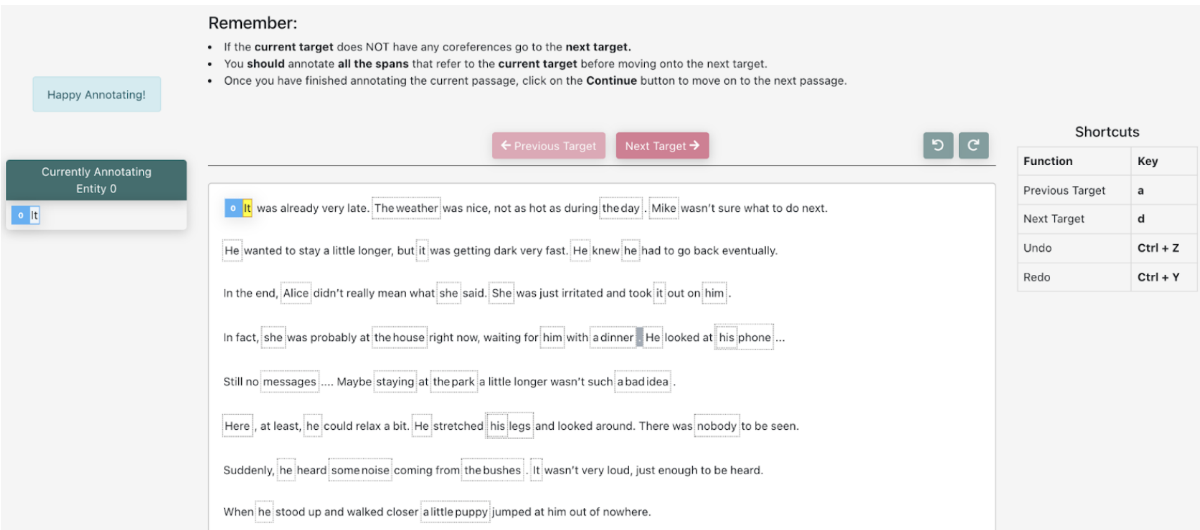
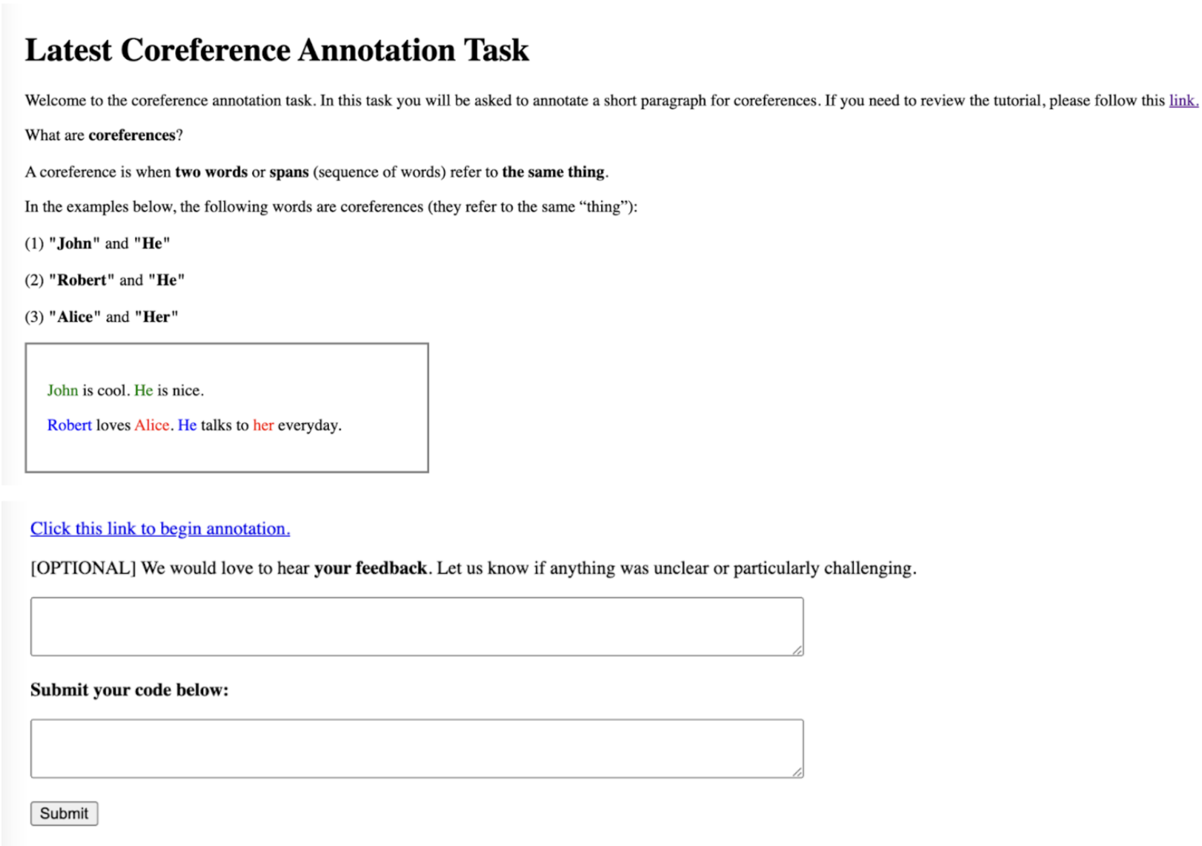


Figure 9: Tutorial interface: quality control example.



Latest 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 10: Annotation task invite on AMT with detailed instructions