Proc. of International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2026)
5-7 February 2026, Boracay Island-Philippines

2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) | 979-8-3315-7191-7/26/$31.00 ©2026 IEEE | DOI: 10.1109/ACDSA67686.2026.11468115

Diverse Word Choices, Same Reference: Annotating Lexically-Rich Cross-Document Coreference

Anastasia Zhukova
University of Göttingen
Göttingen, Germany
0000-0001-9084-2890

Felix Hamborg
Humboldt University
Berlin, Germany
0000-0003-2444-8056

Karsten Donnay
University of Zurich
Zurich, Switzerland
0000-0002-9080-6539

Norman Meuschke
University of Göttingen
Göttingen, Germany
0000-0003-4648-8198

Bela Gipp
University of Göttingen
Göttingen, Germany
0000-0001-6522-3019

***Abstract*—Cross-document coreference resolution (CDCR) identifies and links mentions of the same entities and events across related documents, enabling content analysis that aggregates information at the level of discourse participants. However, existing datasets primarily focus on event resolution and employ a narrow definition of coreference, which limits their effectiveness in analyzing diverse and polarized news coverage where wording varies widely. This paper proposes a revised CDCR annotation scheme of the NewsWCL50 dataset, treating coreference chains as discourse elements (DEs) and conceptual units of analysis. The approach accommodates both identity and near-identity relations, e.g., by linking "the caravan" - "asylum seekers" - "those contemplating illegal entry", allowing models to capture lexical diversity and framing variation in media discourse, while maintaining the fine-grained annotation of DEs. We reannotate the NewsWCL50 and a subset of ECB+ using a unified codebook and evaluate the new datasets through lexical diversity metrics and a same-head-lemma baseline. The results show that the reannotated datasets align closely, falling between the original ECB+ and NewsWCL50, thereby supporting balanced and discourse-aware CDCR research in the news domain.**

Keywords*—cross-document coreference resolution, lexical diversity, media bias, natural language processing.

I. Introduction

Cross-document coreference resolution (CDCR) is a Natural Language Processing (NLP) task that aims to identify and link mentions of the same entities and events across multiple related documents. By identifying when different lexical expressions refer to the same underlying referent, CDCR enables a richer form of content analysis that aggregates information at the level of discourse participants rather than isolated terms. While the ECB+ dataset [1] remains a widely used benchmark for evaluating CDCR models, it primarily emphasizes event resolution and therefore adopts a relatively narrow definition of coreference, i.e., an event is coreferential when it has the same attributes of actor, location, and time [2]. Hence, ECB+ annotates entities as event-dependent and ignores those mentions outside annotated events, thereby preventing the identification of complete actor coverage in the news [3]. A narrow definition of coreference relations falls short when a CDCR model trained on such a dataset is applied to polarized news articles, i.e., articles reporting on the same event but presenting different perspectives, emphasizing distinct aspects of the story, and employing markedly distinct vocabulary [3].

Applying CDCR in the news domain for content analysis and to detect linguistic patterns in media coverage requires a reconsideration of what it means for two mentions to refer to the same event or entity. Traditional coreference models are designed to capture strict referential equivalence (e.g., "Angela Merkel" - "the German chancellor"), yet journalistic discourse often employs "looser" coreference relations, e.g., near-identity [4], quasi-identity [5], euphemisms [6], metaphors [7], [8], or paraphrases [9] that blur referential boundaries [10]. To address this problem, Hamborg et al. [11] proposed NewsWCL50, a dataset that annotated concepts prone to bias by word choice and labeled terms such as "migrants", "caravan", "threat", and "asylum seekers" as referring to the same social group. Capturing the wide variety of word choices provided a more realistic setting for CDCR models and enabled the analysis of how referents are framed, evaluated, and transformed across contexts [12].

While NewsWCL50 captures bias through word choice and labeling instances, the annotation scheme only identifies comparably broad concepts more suitable for content analysis than coreference resolution [13]. This paper proposes a revised annotation scheme for NewsWCL50, which, first, explicitly treats coreference chains as discourse elements (DEs) and conceptual units of analysis, and second, adheres to the definitions and requirements of identity and near-identity relations, providing more fine-grained entities, events, and concepts. We evaluate the annotation scheme by reannotating the entire NewsWCL50 and a subset of ECB+ using the same codebook, and comparing these new datasets using lexical diversity metrics and performance on the same-head-lemma baseline. Our experiments show that the reannotated datasets NewsWCL50$_r$ and ECB+$_r$ exhibit consistently similar metric profiles that lie between those of the original NewsWCL50 and ECB+ datasets. This effectively balances the two codebooks, providing annotations for coreference chains that adhere to conventional definitions of coreference while also addressing the necessity for high lexical diversity in coreferential mentions. The codebook, annotation files in

979-8-3315-7191-7/26$31.00 ©2026 IEEE

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.

MAXQDA, and the final annotations are available at https://github.com/anastasia-zhukova/NewsWCL50r.

II. Related work

Developing datasets for coreference resolution that capture varying levels of complexity and lexical diversity has been a long-standing research focus in within-document coreference resolution, for example, in [5], [14]–[20]. However, annotation schemes and datasets that systematically explore semantic relations and lexical diversity in CDCR remain comparatively rare.

Ahmed et al. [21] addressed the limited lexical diversity of the ECB+ dataset by generating metaphoric paraphrases for event triggers using GPT-4. Although their approach was applied exclusively to events, it demonstrated that introducing paraphrastic variation can effectively increase lexical diversity and, consequently, the complexity of the CDCR task.

As part of the experiments, we extend this idea by reannotating ECB+ to show that lexical diversity can be enriched not only in events but also in entities. We demonstrate that the diverse lexical choices used to refer to a given entity are often present in existing news texts and tend to depend on how an article seeks to frame or portray that entity. However, capturing those diverse entity mentions requires lowering the degree of coreference identity between mentions.

III. Lexically-Rich CDCR Annotation

We propose a reannotation scheme that addresses two key limitations of existing datasets: the overly strict identity relations in ECB+ and the overly broad annotation guidelines in NewsWCL50, which result in discourse elements (DEs) that are either too narrowly or too broadly defined. Our approach integrates and refines annotation rules from both schemes, resulting in a framework that is both concept-centric and fine-grained in nature. This section introduces the core terminology and guiding principles of the proposed annotation scheme. The accompanying codebook in the project repository provides detailed instructions and practical annotation examples.

Mention: We define *a mention* in a text as a *single word* or *multi-word phrase*, i.e., heads and the head modifiers such as adjectives, compounds, adverbs, direct objects, etc., that trigger a coreference chain. Examples: "an anxious and uncertain President Trump," "a difficult emotional decision for the president," "less-educated, native-born Americans," "a cruel action." We annotate only named and nominal mentions [18], and unlike ECB+, we annotate a full noun phrase following the maximum span annotation principle [2].

Coreference chain as a discourse element (DE): We define *a coreference chain* in the linguistic sense as a single, language-independent, semantic unit to be annotated from a set of related articles, i.e., a discourse element. In a document, a chain is represented by a term or a collection of coreferential terms. For example, "the current President of the United States," which can also be referred to as "Trump," "President Trump," "the President," "Donald Trump," etc. In this study, coreferential chains are usually: (1) *entities*, e.g., actors, organizations, geo-political entities (GPEs), locations, objects [14], [18]; (2) *events*, such as short-term actions "deport" or "overstep", or actions or processes with a longer duration, e.g., a meeting, a war, or immigration [19]; (3) *concepts*, i.e., frequently covered yet more broadly defined as story elements that aim to aggregate more abstractly related mentions, e.g., a reaction to one or more events, other consequences thereof, or chains of small but interrelated events [13].

Coreference relations: To capture mentions referring to concepts that vary in semantic complexity and lexical diversity, we define several types of coreferential relations. Each concept is composed of mentions linked by one or more of these relation types. Importantly, every concept must represent a semantically independent element within the news story, ensuring that it can stand as a distinct unit of meaning in the broader discourse.

Identity relations: Identity relations denote a clear equivalence between two mentions that refer to the same concept, such as "Donald Trump" and "the President" [14]. These relations may occur between noun phrases (NPs) and verb phrases (VPs), for example, "met with Donald Trump" and "the planned meeting with Donald." Synonyms grounded in common knowledge or contextual interpretation also constitute identity relations, such as "talked about" – "discussed," "DACA recipients" – "undocumented children," or "Olympics" – "sport competition." While appositional modifiers like "John, an artist" are typically annotated as attributes [14] or less strict identity relations [15], in the context of related news articles, such modifiers should be annotated together with their heads, consistent with other modifiers, for instance, the adjectival modifier in "undocumented children."

Near-identity or bridging relations: To capture relations beyond strict coreference, we adopt the concept of near-identity relations proposed by [4], quasi-identity from [5], and bridging relations as defined by [15]. While such relations are typically omitted in annotation schemes like OntoNotes [14] or ECB+ [1], we include several types of semantically related but non-identical connections between mentions. (1) Part–whole (metonymic or meronymic) relations, where a mention represents a part of a larger concept, e.g., "the Kremlin" – "the Russian government" [4], [15]. (2) Context-dependent equivalences between words with distinct literal meanings that refer to the same concept in context, often conveying evaluative judgment, e.g., "invade" – "cross the border" [11]. These equivalences also include euphemisms [6] and metaphors [7]. (3) Copular constructions formed by verbs such as "be," "seem," or "feel," which associate subjects with attributes, e.g., "This meeting is a big step forward," where "is" links "meeting" and "a big step forward" [22], [23]. (4) Labeling or 'calling' relations, introduced by verbs like "call," "name," "describe," or "denounce," which assign evaluative or ideological labels [24]. For example, in "Trump called Kim Jong Un a Rocket Man," the pair "Kim Jong Un" – "a Rocket Man" forms a bridging relation. Similarly, in "Khan said that Trump behaved like a 12-year-old," the pair "Trump" – "a 12-year-old" should be annotated.

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.

TABLE I.
THE ANNOTATION EXAMPLE DEMONSTRATING THE DIFFERENCE BETWEEN THE ORIGINAL ANNOTATION SCHEMES OF NEWSWCL50 AND ECB+ TO THE PROPOSED LEXICALLY DIVERSE YET FINE-GRAINED ANNOTATION SCHEME.

		Annotations
COUNTRY: USA	**NewsWCL50**	United States; U.S.; White House; CIA director; Matthew Pottinger; Larry Kudlow; chief White House correspondent; American officials; Washington; personal lawyer; C.I.A.; senior director for Asia at the National Security Council; C.I.A. director; U.S. Senators; CIA director and secretary of State-designate; Senate Foreign Relations Committee; director of the East Asia nonproliferation program at the Middlebury Institute; southern White House Director of Trump's National Economic Council; U.S. officials; F.B.I. director; United Nations Command; His administration; Robert Mueller; Jeffrey Lewis; senators; Kudlow; senior vice president at the Center for Strategic and International Studies; Center for Strategic & International Studies; Special Counsel; Central Intelligence Agency Director; U.S. diplomats; American troops; U.S. secretary of state; United States government; U.S.-led forces; Mr. Trump's chief economic adviser; CIA chief; National Security Council; Americans; America; Michael D. Cohen; Mr. Kudlow
	NewsWCL50$_r$	*USA*: United States; U.S.; White House; his administration; Washington; the U.S.-led forces in the conflict; United Nations Command; the White House press office; the Senate; the U.S.; the Senate Foreign Relations Committee; the Washington-based Center for Strategic & International Studies; the State Department; C.I.A.; signers to the armistice; the government; United States government; Senate Foreign Relations Committee; America; the two countries; the U.N. Command
		USA\Kudlow: Mr. Kudlow; Larry Kudlow, Mr. Trump's chief economic adviser; Larry Kudlow, director of Trump's National Economic Council
		USA\Pottinger: Matthew Pottinger, the senior director for Asia at the National Security Council; Matthew Pottinger, senior director for Asian affairs for Trump's National Security Council
		USA\Tillerson: Tillerson; Rex Tillerson
GROUP: Suffered people	**ECB+**	*t37_victims_of_quake*: 24 people; Five; five people; one person; three people
		t37_people_ran_houses: people
		t37_child_killed: 1; child; one; one persona
		t37_50_people_injured: 50; 200; 50 people; dozens; dozens of villagers; five people; hundreds
		t37_2ppl_missing: two other
	ECB+$_r$	14; 140; 1 dead; 10 people; 14 others; 230,000 people; 24 dead; 24 people dead; 249 people injured; 70 others; a child; a child who died when a wall collapsed; A man; an estimated 14 children still trapped under the rubble; Another four people; around 30 people seriously injured; around 50 people with injuries; around 50 people with injuries sustained when the walls of their houses collapsed; at least 24 people; At least five people; at least one person; death toll; dozens; dozens injured; Dozens of people; dozens of villagers; Five dead; Five people; Four other people; hundreds more injured; Injured people; Many people; more than 1,000 people; more than 200; more than 200 people; one man; One of the fatalities; over 200 injured; seven; six children; some 230,000 people around the Indian Ocean; some people; the children still trapped after the mosque collapse in Blang Mancung village; Twelve people; two others missing; two people
ACTOR: Warren Jeffs	**ECB+**	*t36_warren_jeffs*: attorney; FLDS leader's; he; head; him; his; Jeffs; leader; leader Warren Jeffs; pedophile; Polygamist; polygamist leader Warren Jeffs; Polygamist prophet Warren Jeffs; polygamist sect leader Warren Jeffs; Polygamist Warren Jeffs; Warren Jeffs; Warren Jeffs, Polygamist Leader; who
	ECB+r	a handful from day one; a problem; a victim of religious persecution; an accomplice for his role; an accomplice to rape by performing a marriage involving an underage girl; an accomplice to sexual conduct with a minor; an accomplice to sexual misconduct with minors; an accomplice to the rape of a 14-year-old girl; FLDS prophet Warren Jeffs; God's spokesman on earth; her father; his client; Jeffs; Jeffs, 54; Jeffs, who acted as his own attorney; Jeffs, who was indicted more than two years ago; Mr. Jeffs; one individual, Warren Steed Jeffs; one of the most wicked men on the face of the earth since the days of Father Adam; penitent; Polygamist prophet Warren Jeffs; polygamist sect leader Warren Jeffs; polygamist Warren Jeffs; president; prophet of the Fundamentalist Church of the Jesus Christ of the Latter Day Saints; prophet Warren Jeffs; stone-faced; The 54-year-old Jeffs; the defendant; the ecclesiastical head of the Fundamentalist Church of Jesus Christ of Latter Day Saints; the father of a 15-year-old FLSD member 's child; the highest-profile defendant; the prophet; the self-styled prophet; their client; their spiritual leader; This individual; Warren Jeffs; Warren Jeffs, leader of the Fundamentalist Church of Jesus Christ of Latter Day Saints; Warren Jeffs, polygamist leader

Aggregating relations: A single article or a collection of related articles may include mentions connected through structural relations such as set–subset–element or whole–part relations [15], [25]. A typical example is the set–element relation, as in "three women" – "one of these women." In some cases, the set itself is abstract, implicit, or missing entirely, leaving only the mentions of the elements. For instance, in the Comey memos, i.e., a series of documents describing Comey's interactions with Donald Trump, mentions such as "dinners," "meetings," "encounters," and "conversations with Trump" represent elements of a broader but unexpressed set of concepts, namely "Interactions with Trump." Concepts composed solely of element mentions (whether noun or verb phrases) are generally more difficult to identify, as their recognition requires a higher level of abstraction, particularly when the overarching set is not explicitly stated in the text.

DE types: The proposed annotation scheme distinguishes ten DE types to capture semantic diversity and discourse structure across documents. ACTION refers to an activity performed by an actor or another DE [19], such as "Negotiate about the peace" by Trump and Korean officials. At the same time, ACTOR denotes a person proper noun performing an action [18], for example, "Mike Pompeo." COUNTRY represents a geopolitical entity (GE) or its institutions [18], such as "the United States of America (USA)." ACTOR-G is a DE that represents a group of people associated with a country. If DE is labeled as [COUNTRYNAME]-I, it designates individuals who officially represent a country or organization, e.g., "Korean envoys." In contrast, a GE labeled as [COUNTRYNAME]-MISC identifies passive membership or population mentions, such as "Iranian citizens." EVENT captures ongoing or extended activities [19], like an "armed confrontation," and GROUP refers to collectives acting together [1], [26], such as 'demonstrators." OBJECT designates non-animated yet salient items [1], [26], e.g., "DNC servers," while ORGANIZATION denotes formal non-government entities [18], such as "Wikileaks." Finally, MISC encompasses abstract or aggregating concepts [13], for example, "Denu-

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.

clearization,"ß which unify semantically related mentions that lack an explicit set reference in text.

IV. Experiments

A. Dataset reannotation

We reannotated NewsWCL50 and ECB+ using the proposed annotation scheme to compare the characteristics of the resulting datasets. The objective was to create datasets that, despite their differing topical compositions, produce annotated mentions and DEs with comparable distributions in lexical diversity and dataset complexity for CDCR modeling. For instance, although NewsWCL50 primarily contains political news and ECB+ consists mainly of general or human-interest news, the dataset properties of the reannotated versions should become more balanced compared to the significant discrepancies in the original datasets [2].

The reannotated version of NewsWCL50, referred to as *NewsWCL50$_r$*, aimed to produce more precisely defined coreference chains and minimize annotation ambiguity in the original dataset. First, we divided overly broad concepts into multiple, more specific DEs, ensuring that mentions within a DE share coreferential, meronymic, metonymic, or part–whole relations, and that each entity's mentions belong exclusively to a single DE. Second, we annotated previously missing mentions and expanded phrases by including non-annotated noun or verb modifiers. Finally, we added small or previously missed concepts, including singleton GEs. We have reannotated the entire corpus of NewsWCL50 [1].

The reannotation of the ECB+ subset, referred to as *ECB+$_r$*, aimed to achieve a complementary objective of producing more loosely defined, lexically diverse DEs. Specifically, the process (1) treated the annotation of events and entities independently, prioritizing frequency of occurrence over adherence to the event–attribute framework, and (2) expanded the annotation scope from minimum span to maximum span to capture the full range of lexical variation present in the text. ECB+$_r$ consists of five topics from the original test set, which included one political topic and four non-political topics, allowing us to capture diverse lexical choices across both these domains. Specifically, we selected the "36ecbplus," "37ecbplus," "38ecbplus," "39ecbplus," and "41ecbplus" topics, and within each topic, we annotated five articles of comparable length to those in NewsWCL50.

To validate the proposed annotation scheme, we manually curated a randomly selected subset of annotated DEs, i.e., one concept per DE type [28]. The precision of the manual curation for NewsWCL50$_r$ and ECB+$_r$ was close to 0.98. Although complete inter-coder agreement would have strengthened the evaluation, the detailed comparison data analysis of both reannotated datasets below provides a robust complementary validation.

Table I presents representative examples for both datasets. While NewsWCL50$_r$ yielded more clearly defined DEs (e.g., "USA" transformed from one concept to four entities), ECB+$_r$ resulted in more broadly defined DEs compared to the original strict coreferential chains (e.g., "Suffered people") or annotated with more lexical diversity and free from the actor-attribute role (e.g., "Warren Jeffs").

[1]To ensure that NewsWCL50$_r$ can serve as a reliable dataset for validating and testing CDCR models (e.g., [27]), we designate topics 0–3 for validation and topics 4–9 for testing.

B. Data analysis

The data analysis comprises three components: (1) a comparison of the general statistics of the datasets, (2) an analysis of lexical diversity, and (3) a performance evaluation using a simple CDCR baseline. The objective is to compare the original and reannotated datasets and, additionally, to evaluate ECB+$_r$ against the ECB+METAm baseline [21], which enhances lexical diversity in event mentions through GPT-4–based paraphrasing. The statistics presented in this section are calculated based on the six topics/30 documents of the NewsWCL50 test set, as well as the five subtopics/25 documents of ECB+. Consistent with [3], we report the statistics for the original NewsWCL50 excluding the ambiguous DE type ACTOR-I. The following data analysis uses the terms "coreference chain" and DEs interchangeably.

General statistics were assessed based on the number of identified DEs and the number of mentions comprising these DEs. As presented in Table II, NewsWCL50$_r$ exhibits a threefold increase in the total number of DEs, accompanied by a 2.5-fold reduction in average chain size, indicating finer-grained and more precise annotation. The total number of mentions increased by 27.7%, as did the average number of annotated mentions per document, reflecting improved annotation coverage. In contrast, ECB+$_r$ demonstrates an opposite trend, with the number of DEs decreasing by 1.8 times but the average chain size expanding by 6.1 times. Notably, the mean DE size remains comparable between the two datasets, ranging from 14.7 to 15.1 mentions per chain. Although the total number of annotated mentions in ECB+$_r$ increased by 3.5 times, the per-document average remains lower than in NewsWCL50$_r$, mainly due to the shorter length of ECB+ articles.

Lexical diversity is a defining characteristic of CDCR datasets, indicating the extent to which coreferential mentions display paraphrastic variation and semantic flexibility. It is measured using three metrics: (1) the average number of unique head lemmas per coreference chain (UL) [29], (2) the phrasing diversity metric (PD) [2], and (3) the measure of textual lexical diversity (MTLD) [21], [30]. As shown in Table II, NewsWCL50$_r$ and ECB+$_r$ exhibit comparable values for UL and PD. The UL value decreases for NewsWCL50$_r$, reflecting finer-grained DE annotation and reduced ambiguity. At the same time, both UL and PD increase for ECB+$_r$ due to the inclusion of more loosely coreferent mentions. Moreover, ECB+$_r$ demonstrates a substantial increase in MTLD compared to the ECB+METAm baseline, suggesting that higher lexical diversity among annotated mentions can be achieved from the same non-generated text just by using a different annotation scheme. Figure 1 presents the distributions of PD and MTLD, showing that both reannotated datasets follow

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.

TABLE II.
GENERAL STATISTICS AND LEXICAL DIVERSITY METRICS FOR THE ORIGINAL AND REANNOTATED VERSIONS OF NEWSWCL50 AND ECB+. AN ASTERISK (*) DENOTES THAT A SUBSET OF FIVE SUBTOPICS WAS USED.

Dataset	Coreference chains / Discourse elements (DEs)					Mentions				Lexical diversity		
	all	entity	events	singletons	avg. size	all	entity	event	avg. per doc	UL	PD	MTLD
NewsWCL50	134	96	38	4	38.2	5115	3886	1229	102.3	10.71	8.99	14.54
NewsWCL50$_r$	433	374	59	102	15.1	6531	4758	1773	130.6	6.05	9.06	15.58
ECB+*	171	112	59	59	2.4	407	246	161	16.3	2.19	1.99	4.55
ECB+METAm*	168	109	59	57	2.4	399	240	159	16.0	2.92	3.25	6.82
ECB+$_r$*	97	84	13	25	14.7	1427	958	469	57.1	5.88	9.04	20.65

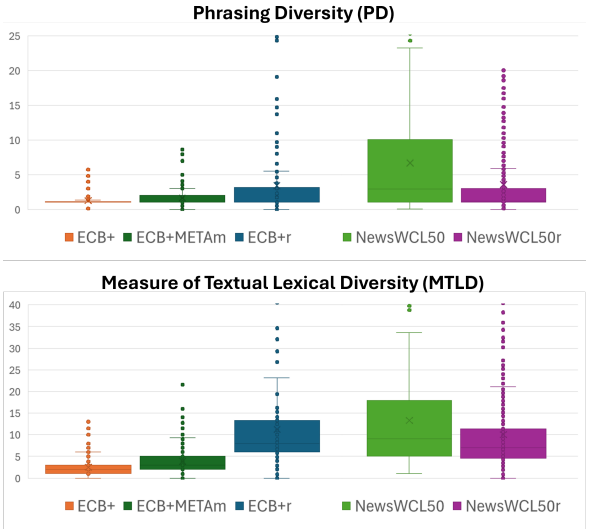

Fig. 1. Distribution of lexical diversity measured by PD and MTLD. Both NewsWCL50$_r$ and ECB+$_r$ exhibit comparable distributions, demonstrating that the proposed annotation scheme achieves a balanced level of lexical diversity in the coreference chains or DEs across news articles, regardless of domain differences.

similar patterns. This consistency suggests that the proposed annotation scheme yields more balanced coreference chains, thereby avoiding the overly broad annotations of the original NewsWCL50 and the excessively narrow ones of ECB+.

Performance analysis was conducted using a simple same-head-lemma baseline [1]. Model performance was evaluated with the CoNLL F1 score, i.e., a standard metric for coreference resolution that averages the B3, MUC, and CEAF$_e$ measures (e.g., as applied in [1], [27], [29]). We evaluated non-singleton chains only. As shown in Table III, NewsWCL50$_r$ and ECB+$_r$ achieve comparable results (54.08 vs. 52.92), representing a more balanced performance relative to the larger discrepancies observed in the original datasets. This outcome indicates that the reannotated datasets provide a moderate level of difficulty for CDCR models, i.e., neither as easily resolvable as the original ECB+ chains nor as challenging due to excessive semantic breadth as in the original NewsWCL50.

The general statistics (Table II) demonstrate consistent annotation patterns across datasets, while lexical diversity metrics confirm that both corpora achieve balanced variation in wording and semantic richness. Furthermore, comparable model performance scores in the baseline (Table III) indicate that the reannotated datasets yield a moderate and comparable

TABLE III.
THE PERFORMANCE OF THE SAME-HEAD-LEMMA BASELINE.

Dataset	MUC	B3	CEAF$_e$	CoNLL
NewsWCL50	82.49	49.70	11.96	48.05
NewsWCL50$_r$	79.59	51.82	30.82	54.08
ECB+*	68.35	73.27	68.10	69.91
ECB+METAm*	49.06	67.22	57.85	58.04
ECB+$_r$*	82.32	47.98	28.46	52.92

level of difficulty for CDCR models. Together, these findings validate the proposed scheme as both methodologically sound and effective in producing lexically diverse, semantically coherent annotations that are suitable for both NLP and social science research.

C. Discussion

The proposed annotation scheme focuses on a lexical diversity challenge for CDCR models, requiring them to learn and recognize looser coreference relations that link phrases with diverse word choices [2]. By incorporating paraphrases, metonymic relations, euphemisms, metaphors, and evaluative wording, the dataset compels models to move towards capturing deeper semantic and contextual equivalences. This not only tests model robustness but also encourages the development of CDCR systems capable of handling the linguistic variability typical of real-world news discourse, particularly in the study of media bias, framing, and discourse [3]. This capability bridges computational methods with critical approaches in media and communication studies, enabling large-scale content analysis that captures not only what is being discussed, but also how and why it is framed in particular ways [12].

V. CONCLUSION

The proposed annotation scheme advances CDCR by integrating lexical diversity and looser identity relations into a consistent and balanced annotation framework. By refining event and entity boundaries while preserving semantic variability, it produces datasets that better reflect the linguistic complexity of real-world news discourse. This approach not only presents a meaningful challenge for CDCR models, requiring them to recognize paraphrastic and context-dependent relations, but also enables large-scale, data-driven analyses of media bias, framing, and discourse.

REFERENCES

[1] A. Cybulska and P. Vossen, "Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.

(LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 4545–4552. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/840{_}Paper.pdf

[2] A. Zhukova, F. Hamborg, and B. Gipp, "Towards evaluation of cross-document coreference resolution models using datasets with diverse annotation schemes," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4884–4893. [Online]. Available: https://aclanthology.org/2022.lrec-1.522/

[3] A. Zhukova, F. Hamborg, K. Donnay, and B. Gipp, "XCoref: Cross-document coreference resolution in the wild," in *Information for a Better World: Shaping the Global Future*, M. Smits, Ed. Cham: Springer International Publishing, 2022, pp. 272–291. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-96957-8_25

[4] M. Recasens, E. Hovy, and M. A. Martí, "A typology of near-identity relations for coreference (NIDENT)," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: https://aclanthology.org/L10-1103/

[5] E. Hovy, T. Mitamura, F. Verdejo, J. Araki, and A. Philpot, "Events are not simple: Identity, non-identity, and quasi-identity," in *Workshop on Events: Definition, Detection, Coreference, and Representation*, E. Hovy, T. Mitamura, and M. Palmer, Eds. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 21–28. [Online]. Available: https://aclanthology.org/W13-1203/

[6] M. Gavidia, P. Lee, A. Feldman, and J. Peng, "CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 2658–2671. [Online]. Available: https://aclanthology.org/2022.lrec-1.285/

[7] R. Joseph, T. Liu, A. B. Ng, S. See, and S. Rai, "NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 090–10 104. [Online]. Available: https://aclanthology.org/2023.findings-acl.641/

[8] B. Radford, "Seeing the forest and the trees: Detection and cross-document coreference resolution of militarized interstate disputes," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, A. Hürriyetoğlu, E. Yörük, V. Zavarella, and H. Tanev, Eds. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 35–41. [Online]. Available: https://aclanthology.org/2020.aespen-1.7/

[9] J. P. Wahle, B. Gipp, and T. Ruas, "Paraphrase types for generation and detection," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 148–12 164. [Online]. Available: https://aclanthology.org/2023.emnlp-main.746/

[10] T. Heuss, B. Humm, C. Henninger, and T. Rippl, "A comparison of ner tools wrt a domain-specific vocabulary," in *Proceedings of the 10th International Conference on Semantic Systems*, 2014, pp. 100–107.

[11] F. Hamborg, A. Zhukova, and B. Gipp, "Automated identification of media bias by word choice and labeling in news articles," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Jun. 2019.

[12] F. Hamborg and K. Donnay, "NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 1663–1675. [Online]. Available: https://aclanthology.org/2021.eacl-main.142/

[13] M. Schreier, *Qualitative Content Analysis in Practice*. SAGE Publications Ltd, 2012.

[14] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, A. Taylor, C. Greenberg, E. Hovy, R. Belvin *et al.*, "Ontonotes release 4.0," *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 2011.

[15] T. O'Gorman, K. Wright-Bettner, and M. Palmer, "Richer event description: Integrating event coreference with temporal, causal and bridging annotation," in *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 47–56. [Online]. Available: https://www.aclweb.org/anthology/W16-5706

[16] T. Mitamura, Z. Liu, and E. H. Hovy, "Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track." in *TAC*, 2017.

[17] ——, "Overview of tac kbp 2015 event nugget track." in *TAC*, 2015.

[18] L. D. Consortium *et al.*, "Ace english annotation guidelines for entities," Technical report, Linguistic Data Consortium, Tech. Rep., 2008.

[19] ——, "Ace (automatic content extraction) english annotation guidelines for events. version 5.4. 3," *ACE*, 2005.

[20] M. Recasens, M. A. Martí, and C. Orasan, "Annotating near-identity from coreference disagreements," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 165–172. [Online]. Available: https://aclanthology.org/L12-1391/

[21] S. R. Ahmed, Z. E. Wang, G. A. Baker, K. Stowe, and J. H. Martin, "Generating harder cross-document event coreference resolution datasets using metaphoric paraphrasing," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 276–286. [Online]. Available: https://aclanthology.org/2024.acl-short.27/

[22] D. Wedding, "The social psychology of stigma, edited by todd f. heatherton, robert e. kleck, michelle r. hebl and jay g. hull," *JOURNAL OF PSYCHIATRY AND LAW*, vol. 30, no. 1, pp. 99–100, 2002.

[23] P. Cap, "Metaphor and political discourse: Analogical reasoning in debates about europe: by andreas musolff, palgrave macmillan, houndmills, basingstoke, 2004, viii+ 211 pp., hb," 2006.

[24] S. Kosloff, J. Greenberg, T. Schmader, M. Dechesne, and D. Weise, "Smearing the opposition: Implicit and explicit stigmatization of the 2008 us presidential candidates and the current us president." *Journal of Experimental Psychology: General*, vol. 139, no. 3, p. 383, 2010.

[25] H. H. Clark, "Bridging," in *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing, TINLAP 1975, Stroudsburg, PA, USA. Association for Computational Linguistics*, 1975, pp. 169—174.

[26] A. Cybulska and P. Vossen, "Semantic relations between events and their time, locations and participants for event coreference resolution," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, R. Mitkov, G. Angelova, and K. Bontcheva, Eds. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2013, pp. 156–163. [Online]. Available: https://aclanthology.org/R13-1021/

[27] M. Bugert and I. Gurevych, "Event coreference data (almost) for free: Mining hyperlinks from online news," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 471–491. [Online]. Available: https://aclanthology.org/2021.emnlp-main.38/

[28] R. Steinberger, S. Hegele, H. Tanev, and L. Della Rocca, "Large-scale news entity sentiment analysis." in *RANLP*, 2017, pp. 707–715.

[29] A. Eirew, A. Cattan, and I. Dagan, "WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 2498–2510. [Online]. Available: https://aclanthology.org/2021.naacl-main.198/

[30] P. M. McCarthy and S. Jarvis, "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior Research Methods*, vol. 42, no. 2, pp. 381–392, 2010. [Online]. Available: https://doi.org/10.3758/BRM.42.2.381

Authorized licensed use limited to: Universitaet Goettingen. Downloaded on November 04,2026 at 15:38:56 UTC from IEEE Xplore. Restrictions apply.