GOLEMcoref: A Multilingual Coreference Dataset of Fiction

Andreas van Cranenburgh,[*1] Xiaoyan Yang,[*1] Alvanita,[2] Cecilia Nicole Di Domenico,[3] Maria Ferragud,[4] Arianna Graciotti,[1] Byungjun Kim,[5] Seonyeong Park,[5] Noa Visser Solissa,[1] Xiaoyu Zhou,[1] Federico Pianzola[1]

[1]CLCG, University of Groningen, [2]Coventry University, [3]University of Trento, [4]Universitat Jaume I de Castellón, [5]The Academy of Korean Studies

Correspondence: f.pianzola@rug.nl

Abstract

We present a multilingual coreference dataset of 827k tokens of fiction in 7 languages: Bahasa Indonesia, Chinese, Dutch, English, Italian, Korean, and Spanish. The dataset includes full stories of diverse lengths, ranging from 500 to 17k words. We discuss our annotation scheme focusing on characters and language-specific challenges we encountered. Finally we present evaluation results of a neural coreference system trained on our dataset. We show that jointly training a system across all languages provides a strong improvement over monolingually trained models. The dataset is available under a creative commons license in CoNLL-2012 and CorefUD format at https://github.com/GOLEM-lab/GOLEMcoref/

1 Introduction

Coreference resolution remains a challenging task in natural language processing, since it depends on decisions that transcend sentence boundaries requiring full discourse context and world knowledge. Despite considerable progress, most research still focuses predominantly on English. In addition, fictional texts bring additional challenges not covered by standard benchmark datasets, such as long documents and a large proportion of pronouns and dialogue (Krug et al., 2015). The type and length of texts in commonly used datasets is predominantly short news texts (e.g., OntoNotes, 500 words on average) or older literature (e.g., LitBank, 2000 word fragments). Therefore, new data resources and models are needed for cultural analytics.

We address these challenges by presenting an annotated multilingual dataset of full-text contemporary fanfiction, both short and long. This makes it possible to get a complete picture of character arcs and their contribution to the narrative. Since fanfiction is a form of user-generated content, it tends to be quite different from the carefully edited language in published texts. We present the first coreference dataset of fiction for Bahasa Indonesia, Italian, and Spanish.

2 Background

The task of coreference resolution consists of identifying mentions in a text that refer to the same entity. Our work addresses three shortcomings of current coreference datasets: (1) multilinguality, (2) lack of genre diversity, and (3) short texts.

The CoNLL-2012 shared task (Pradhan et al., 2012) was a milestone for coreference resolution and the OntoNotes dataset (Weischedel et al., 2013) used in it has been established as the standard benchmark for coreference resolution. Throughout the years, improvements in performance have been made: from 63.4 for the best English system in the shared task to the recent 83.6 (Martinelli et al., 2024). However, while OntoNotes is a multi-genre and multilingual dataset, it predominantly contains news text, and most attention in NLP has gone towards English.

The CorefUD (Nedoluzhko et al., 2022) initiative is an effort to bundle coreference annotations for different languages in a common format, inspired by the Universal Dependencies initiative. It has been used in several shared tasks that aim to address the challenges of multilingual coreference.

Several languages already have annotated coreference datasets of fiction texts. LitBank (Bamman et al., 2020) consist of 2000 word fragments of English novels. OpenBoek (van Cranenburgh and van Noord, 2022) contains classic Dutch literature, providing 10,000 word fragments of novels. KoCoNovel (Kim et al., 2024) is a dataset of 50 full-length Korean short stories, 3,500 tokens on average. NovelCR is a parallel Chinese-English dataset of webnovels, designed for long-span coreference resolution (Tong and Wang, 2025). A few

[*]Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 472–480

more datasets exist for languages not covered in our corpus; e.g., French (Mélanie-Becquet et al., 2024) and German (Krug et al., 2018; Pagel and Reiter, 2020).

Most coreference datasets either contain short texts, or divide longer texts into shorter fragments. However, long document coreference has been recognized as an important challenge (Toshniwal et al., 2020; Martinelli et al., 2025a,b). Another challenge, especially for non-English coreference, are zero anaphora, where an anaphoric relation is expressed by an implied mention, e.g., through a verb inflection in a pro-drop language. Proper handling of zero anaphora involves predicting empty nodes in a separate stage (Straka, 2023), or jointly predicting empty nodes and coreference (Chen et al., 2021; Straka, 2024, 2025).

3 Corpus

Our corpus was compiled from three online fiction platforms: Archive of Our Own (AO3), Postype, and Wattpad. These platforms provide large quantities of narrative fiction suitable for studying coreference phenomena. We collected texts in seven languages (Bahasa Indonesia, Chinese, Dutch, English, Italian, Korean, and Spanish), targeting approximately 100k words per language. Fanfiction stories reuse characters and other elements of known narrative universes. To maintain a balanced and diverse corpus across languages, we sampled short stories ranging from 200 to 17,000 words and diversified our selection along multiple dimensions, including fandom (e.g., Star Wars, BTS), settings (e.g., fantasy, college), narrative perspective (e.g., first-person POV), dialogue–narration ratio, and themes (e.g., jealousy, war). See Appendix A for all selection criteria.

3.1 Annotation Scheme

We developed language-specific annotation instructions through weekly discussions with annotators to ensure practical applicability. Our annotations are restricted to characters, which include all animate mentions; all inanimate mentions are excluded. This is in contrast with other coreference datasets such as OntoNotes (unrestricted, all NPs are markables), and LitBank (entities in specific ACE categories are mentions), which also annotate places, things, and ideas as mentions. The annotation includes singletons (mentions that occur only once and therefore do not corefer). The guidelines specify a typology of character mentions. These include proper names, pronouns (personal, possessive, demonstrative, reflexive, and indefinite), specific noun phrases, generics, combinations of generics and proper names, periphrases, and metonymic expressions that clearly refer to a character. A notable inclusion is zero anaphora, which are prevalent in pro-drop languages such as Chinese, Italian, and Spanish. For example, in the Spanish sentence "Cerró la puerta" ([He] closed the door), the omitted subject is annotated using a zero-width span. In Bahasa Indonesia, a possessive pronoun can be a merged zero pronoun; e.g., in "Dia membaca buku*nya* sambil minum kopi" (He read book[his] while drinking coffee), the suffix *-nya* (his/her), refers back to the subject *Dia* (He). Whenever a mention is implied or expressed as part of a word, an empty node is added after the respective word. This addresses a gap in prior resources and enhances cross-lingual consistency in coreference chains (Iida and Poesio, 2011; Aloraini and Poesio, 2020; Chen et al., 2021).

As coreference links we consider not only identity relations, but also predicative and appositive relations. When plural entities consisting of previously mentioned characters are mentioned, e.g., "John and Mary" and "they" or "the couple," we annotate this with an *includes* relation (also known as a split antecedent). There are also instructions for complex mention types, such as noun phrases that include modifying clauses: "the man in the room" is a single mention span. Furthermore, mentions in negated or hypothetical contexts are only annotated when the character exists in the narrative world. For more details see the full guidelines in the repository.

3.2 Annotation Procedure

For each language there were at least 2 native-speakers annotators who annotated all documents for that language, using INCEpTION (Klie et al., 2018). Annotators were Master's students or persons with a Master's degree living in the Netherlands, who have been paid or received course credits based on the number of hours worked. Annotators had considerably different paces, taking from 76 to 151 hours to complete the task. In a separate curation step, a different person (a PhD candidate or expert researcher) per language reviewed the annotations for each document, and decided on an adjudicated version. To keep the quality of annotation as high as possible and avoid cognitive fatigue,

	Chinese	Dutch	English	Indonesian	Italian	Korean	Spanish	*Total*
Tokens	104,772	119,989	134,348	132,719	120,058	90,031	125,086	827,003
Sentences	3,907	9,953	8,836	10,901	6,148	9,686	6,221	55,652
Zero anaphora	68	0	36	4,939	5,831	39	4,703	15,616
Split antecedents	208	670	436	163	231	320	403	2,431
Mentions	8188	16,427	18,511	19,368	14,491	9,975	15,722	102,682
Entities	570	913	741	684	552	667	816	4,943
tokens / sent	26.8	12.1	15.2	12.2	19.5	9.3	20.1	14.9
mentions / tokens	0.078	0.137	0.138	0.146	0.121	0.111	0.126	0.124
mentions / entities	14.36	17.99	24.98	28.32	26.25	14.96	19.27	20.77

Table 1: Dataset statistics. Token counts include punctuation and zero anaphora.

annotators have been instructed to work in short sessions of maximum 2 hours, for a maximum of 8 hours per week. In the first few weeks, several meetings have been held between the annotators and the research team, to discuss challenges in the application of the guidelines (developed in English) to other languages and to discuss difficult or ambiguous examples.

3.3 Language-specific challenges

We encountered several challenges during annotation and curation, which can be categorized into three main domains: cross-linguistic structural issues, narrative and fictional complexity, and language-specific idiosyncrasies. A detailed report with examples is available in the repository.

The most widespread cross-linguistic challenge lies in resolving unclear referents. This is evident in pronoun ambiguity (e.g., the Chinese 她 /her referring ambiguously to one of two female characters), the interpretation of collective mentions (e.g., English "the five of them," Spanish *ellos*/they), and the prevalence of zero anaphora, where omitted subjects or objects must be inferred from context.

Second, complex nominal phrases require annotators to identify the right span (e.g., Chinese and Korean have frequent nested descriptions like "the squad that arrests Muggle-born wizards"; or English "the one who's 'taking care' of Peter"), rather than just marking the head noun.

Third, literary devices, fantasy elements, and narrative strategies introduce non-standard referents. These require distinguishing out-of-storyworld addressees (the reader-directed "you" vs. an in-story character), interpreting supernatural entities (e.g., a speaking wolf), tracking identity transformations (e.g., a character whose body is inhabited by another character), and inference from stylistic devices such as synecdoche (e.g., "eyes" used as a grammatical subject and referring to a character).

Finally, there are various language-specific features. These encompass morphological attachment in agglutinative languages (e.g., Korean case particles, the Indonesian clitic *-nya*) and word segmentation issues in Chinese (你自己 /yourself), and the encoding of multiple implicit arguments within a single Italian verb form (e.g., *spiegandoglielo* [explaining it to him]). Furthermore, cultural-linguistic hybridity, such as honorifics and inconsistent name transliterations (e.g., "Ferenc" and "Franz" for the same character in Korean), often requires fandom knowledge for accurate coreference resolution.

4 Results

We make the annotations available in two formats: CoNLL-2012 because it is expected by most coreference systems, and CorefUD because it supports zero anaphora and split antecedents, which we annotate (see Appendix B). For CorefUD we use Stanza (Qi et al., 2020) to add universal dependencies and POS tags to the texts. Table 1 reports basic statistics of the dataset. A more detailed breakdown by data split is provided in Table 5 (see Appendix C). Zero anaphora are a pervasive element in three of the languages (Bahasa Indonesia, Italian, Spanish), while for the other languages there are only a few (Chinese, English, Korean), or none at all (Dutch).

We evaluate the annotator agreement and coreference systems using the CoNLL metric (Pradhan et al., 2012). In addition we report the mention F1 score which is useful for diagnosing to what extent mention identification holds back the coreference performance. We also report the LEA score (Moosavi and Strube, 2016), which is an alternate coreference metric that addresses shortcomings identified in the CoNLL metric; e.g., giving more weight to larger entities, which is useful when evaluating longer documents. For agreement we also

report the coreference of zero mentions (Zero F1) as reported by the CorefUD scorer (Žabokrtský et al., 2022).

Language	Mention	Zero	CoNLL	LEA
Chinese	94.8	11.4	90.4	90.5
Dutch	91.7	-	77.3	81.4
English	91.9	19.0	77.6	85.8
Indonesian	88.0	59.2	69.4	74.2
Italian	80.2	56.7	63.9	63.0
Korean	87.3	0.0	71.0	74.8
Spanish	79.7	26.2	63.4	59.1

Table 2: Annotator agreement (F1) before curation.

4.1 Annotator Agreement

All documents were independently annotated by at least 2 annotators, so we can evaluate annotator agreement on the full dataset. For each pair of annotators, we calculate the agreement by running coreference evaluation with one annotator as the "gold file," and the other as the "system file"; see Table 2. For each language, all documents are evaluated at once and reported in a single micro-averaged F1 score. This means that longer, harder documents receive more weight, to better show actual annotation difficulty.

The agreement scores give an idea of the difficulty of the annotation task for this dataset, as well as the degree to which the guidelines help the annotators to arrive at the same decisions. Further quality control was done using a check list approach (e.g., mention spans should not start or end with punctuation). The final release of the dataset and the trained coreference models use the curated, check-listed annotations.

It is noteworthy that Chinese achieved the highest agreement among all languages in our study (90.4 CoNLL), considerably above previously reported scores of 81.55 for dialogue transcripts (Zheng et al., 2023) and 65.9 for full-length novels (Jiang et al., 2023). This may partially reflect annotator calibration during the pilot phase, where annotators collaborated to develop shared intuitions about character coreference. Additionally, the high information density of Chinese text allowed annotators to view full sentences without excessive scrolling in INCEpTION, potentially reducing cognitive load. This is also consistent with Chinese having the shortest average annotation time (100 hours) across all languages.

4.2 Evaluation of Coreference Systems

To offer a baseline for future research, we train coreference systems to predict mentions and coreference links (i.e., we do not predict split antecedent relations). We evaluate two systems: `fast-coref` (Toshniwal et al., 2021) and xCoRe (Martinelli et al., 2025b). For both systems, we use mmBERT (Marone et al., 2025) as the document encoder, since it supports all the languages in our dataset. We also experimented with mDeberta, but found mmBERT to give better results. We set the maximum segment length to 8192, the maximum for mmBERT. We use the same parameters as used for LitBank in Toshniwal et al. (2021) and Martinelli et al. (2025b), except that we do not use crossvalidation. While `fast-coref` and xCoRe support long documents, they have no support for predicting zero anaphora. For these experiments we therefore a insert special dummy token `<EMPTY>` in the input for each zero anaphora;[1] i.e., the system does not have to predict where in each sentence zero anaphora occur.[2]

We held out 10% as test data, and 10% as development data, for each language. Each set is balanced to contain both short and longer documents. We first trained separate models for each language, see Table 3. Considering the limited amount of tokens per language (90–134k), the amount of training data is a limiting factor. Therefore we experimented with augmenting the training data with external datasets: LitBank for English, and Openboek for Dutch. We improve the alignment with our annotation scheme by only keeping coreference annotations of person entities, yielding variants LitBank^p and Openboek^p. For English, Table 3 shows improvement. For Dutch there is a decrease, which could be due to annotation scheme differences.

Since the annotation scheme and document embeddings are the same for all languages, we can ex-

[1]The zero anaphora are annotated as zero width spans directly after the word that triggers them. However, this is not always a good position in which to insert the dummy token, since it might interrupt another span. When the position of the zero width span would fall in another span, we search for the closest position to the right that does not fall in any span.

[2]In future work an end-to-end system with proper treatment of zero anaphora should be trained. However, we have not found a coreference system that handles both long documents and zero anaphora. Furthermore, a system that handles zero anaphora (e.g., Straka, 2025) expects syntactic information for the zero anaphora (i.e., whether the zero anaphor is a subject or object, and of which verb), but we have not annotated such information.

	fast-coref			xCoRe		
Language	**Mention F1**	**CoNLL**	**LEA F1**	**Mention F1**	**CoNLL**	**LEA F1**
Chinese	91.7	63.2	52.6	89.1	57.1	51.6
Dutch	96.9	**66.3**	65.8	96.7	54.8	47.0
English	95.2	58.8	56.0	91.9	50.6	41.7
Indonesian	95.0	70.5	66.6	94.4	58.7	53.8
Italian	94.2	67.3	70.1	93.1	64.0	63.1
Korean	**91.1**	63.9	60.5	89.2	58.8	59.6
Spanish	91.6	63.3	55.6	90.1	59.4	48.7
Dutch+OB	97.2	63.8	59.2	97.0	60.3	56.4
English+LB	95.4	62.9	61.0	94.4	53.1	44.1

Table 3: Results for monolingually trained models (including singletons, gold empty nodes); OB and LB refer to Openboek^p and LitBank^p, respectively. Bold indicates best score for language and metric.

	fast-coref			xCoRe		
Language	**Mention F1**	**CoNLL**	**LEA F1**	**Mention F1**	**CoNLL**	**LEA F1**
Chinese	**92.9**	**68.6**	**64.5**	92.2	65.3	60.5
Dutch	97.5	65.7	64.2	**97.8**	64.2	**67.1**
English	95.4	**63.1**	**61.2**	**95.5**	58.1	53.7
Indonesian	**96.7**	**72.6**	**71.6**	**96.7**	61.1	55.5
Italian	**95.0**	**74.8**	**73.9**	94.4	69.6	72.7
Korean	90.9	**66.9**	**67.2**	90.6	63.4	63.7
Spanish	**93.3**	**66.2**	**57.8**	91.7	58.8	46.6

Table 4: Results for a single, crosslingually trained model, augmented with Openboek^p and LitBank^p data (including singletons, gold empty nodes). Bold indicates best score for language and metric.

periment with combining the training data of all the languages and jointly training a single crosslingual model. We also include LitBank^p and Openboek^p in the training data, for a total of 1.1M tokens. The results are in Table 4. We find that joint training substantially improves the results for 6 out of 7 languages, by up to 7 percentage points. This is in line with previous results on crosslingual coreference, e.g., Straka and Straková (2022).

5 Conclusion

We have presented a multilingual dataset of coreference in fiction, including documents of various lengths, with a total of 827k tokens. A multilingual dataset in 7 languages annotated with the same annotation scheme and available in CorefUD format is extremely useful for improving multilingual coreference resolution. We obtained promising results with data augmentation and crosslingual training. Future work should address the handling of zero anaphora and split antecedents. We hope that our dataset will boost research into multilingual coreference resolution, as well as computational literary studies.

Limitations

While all annotations have been double checked, since each text has at least two annotators and a curator, it remains difficult to achieve perfect annotation consistency and quality. The amount of data per language (90–134k tokens) is less than that in commonly used coreference datasets (OntoNotes 1.7M, LitBank 210k). Especially in the monolingual setting, the training data size is a limiting factor. The results we have presented are not for an end-to-end system, since empty nodes for zero anaphora need to be inserted into the input explicitly. Future research should address this, such that a plain text can be given as input. For reasons of time and scope we have only experimented with two coreference systems. However, other coreference system may produce better results, such as Cor-

pipe (Straka, 2025). Finally, all our experiments use a single embedding model as the document encoder (mmBERT). Previous work suggests that language-specific monolingual models can give superior results. However, such models would not be usable in a crosslingual setting.

Ethical considerations

We followed best practices for using online fandom data: obtaining permission, attribution, giving back, and learning community norms (Dym and Fiesler, 2020). People within fan communities often have protective views regarding their data and its use by researchers. One of the risks is that of amplifying fan content to an audience it was never intended for, which could compromise the privacy and context of that data. Fan-generated data are highly contextual to the owner and their specific privacy needs. To address these concerns, all stories included in our dataset are either "orphaned works" for which authors voluntarily waived copyright,[3] or we obtained explicit permission from authors to analyze and share the data after anonymization (see Appendix D).

Acknowledgments

We are grateful to the anonymous reviewers for helpful suggestions. This work is supported by the ERC grant GOLEM (101040938), funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Abdulrahman Aloraini and Massimo Poesio. 2020. Anaphoric zero pronoun identification: A multilingual approach. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 22–32.

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

Shisong Chen, Binbin Gu, Jianfeng Qu, Zhixu Li, An Liu, Lei Zhao, and Zhigang Chen. 2021. Tackling zero pronoun resolution and non-zero coreference resolution jointly. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 518–527.

Brianna Dym and Casey Fiesler. 2020. Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures*, 33.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 804–813.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. KoCoNovel: Annotated dataset of character coreference in Korean novels. *Preprint*, arXiv:2404.01140.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2018. Description of a corpus of character references in German novels - DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers Nr. 27.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394.

Giuliano Martinelli, Tommaso Bonomo, Pere-Lluís Huguet Cabot, and Roberto Navigli. 2025a. BOOKCOREF: Coreference resolution at book scale. In

[3] https://archiveofourown.org/faq/orphaning?language_id=en

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24526–24544.

Giuliano Martinelli, Bruno Gatti, and Roberto Navigli. 2025b. xCoRe: Cross-context coreference resolution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34252–34266.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Frédérique Mélanie-Becquet, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau. 2024. Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. *Journal of Computational Literary Studies*, 3:1–34.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Janis Pagel and Nils Reiter. 2020. GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 55–64.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

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

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106.

Milan Straka. 2025. CorPipe at CRAC 2025: Evaluating multilingual encoders for multilingual coreference resolution. In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 130–139.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37.

MeiHan Tong and Shuai Wang. 2025. NovelCR: A large-scale bilingual dataset tailored for long-span coreference resolution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5161–5173.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526.

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

Andreas van Cranenburgh and Gertjan van Noord. 2022. Openboek: A corpus of literary coreference and entities with an exploration of historical spelling normalization. *Computational Linguistics in the Netherlands Journal*, 12:235–251.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0. Linguistic Data Consortium.

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

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

A Criteria for Corpus Construction

1. For each language, identify works on AO3 for which copyright has been waived.
2. Use metadata (AO3 Warnings) and final manual check to exclude works with explicit sexual content, to maintain the dataset's ethical integrity and appropriate content scope.
3. Remove works shorter than 200 words or longer than 13,000 words, to exclude extremely brief or overly long texts that may

distort the overall word balance and reduce annotation feasibility within the planned time frame.

4. Include at least three works with lengths around 10,000 words, sampling them from three different fandoms, to ensure sufficient coverage of long-form narratives and diverse character interactions.
5. Ensure that some works are under 1,000 words, to capture short-form narratives and highly condensed reference phenomena.
6. Diversify the selection among fandoms as much as possible, aiming for a variety of story worlds, characters, and narrative settings.
7. Include a mix of dialogue-heavy and narrative-heavy works, as well as texts written from different narrative perspectives, to reflect varied storytelling styles and pronoun usage patterns.
8. Maintain a mix of popular and less popular works (kudos/hits ratio), to ensure that the corpus reflects both mainstream and niche writing styles, as well as good and poor writing.
9. Diversify additional narrative features by leveraging author-supplied Additional Tags, which describe elements such as plot events (e.g., First Kiss), themes (e.g., Betrayal), tone (e.g., Angst), style (e.g., Stream of Consciousness), and narrative devices (e.g., Time Loop).
10. If the number of remaining works is not enough to reach 100k words in total, or it is not varied enough, look for additional works that fill the criteria and contact the authors to get consent.
11. If not enough works are available on AO3, look for more works on other online reading platforms (e.g., Wattpad, Postype).

B Example annotations

See Figure 1 and Figure 2 for examples of annotations in the CoNLL 2012 and Coref UD format. Both formats use the last column for coreference information. The former uses dummy tokens (`<EMPTY>`) for zero anaphora. The latter has actual empty nodes for zero anaphora, see the line with ID 4.1 in Figure 2. In addition, the `SplitAnte` attribute specifies that the "we" entity e4 is composed of entities e0 and e3. For readability the morphological features have been abbreviated.

Language	Split	Stories	Tokens	Mean	Median
Chinese	Dev	2	6,113	3,056.5	3,056.5
Chinese	Test	2	6,452	3,226.0	3,226.0
Chinese	Train	20	92,207	4,610.4	3,101.0
Chinese	All	24	104,772	4,365.5	3,101.0
Dutch	Dev	3	14,546	4,848.7	3,322.0
Dutch	Test	3	15,386	5,128.7	3,753.0
Dutch	Train	20	90,057	4,502.9	2,761.5
Dutch	All	26	119,989	4,615.0	3,040.5
English	Dev	3	13,665	4,555.0	2,729.0
English	Test	3	13,547	4,515.7	2,795.0
English	Train	24	107,136	4,464.0	2,583.0
English	All	30	134,348	4,478.3	2,657.0
Indonesian	Dev	4	8,547	2,136.8	1,999.0
Indonesian	Test	4	9,436	2,359.0	2,091.0
Indonesian	Train	34	114,736	3,374.6	2,219.0
Indonesian	All	42	132,719	3,160.0	2,219.0
Italian	Dev	3	5,972	1,990.7	1,892.0
Italian	Test	3	7,274	2,424.7	2,301.0
Italian	Train	28	106,812	3,814.7	2,461.0
Italian	All	34	120,058	3,531.1	2,443.5
Korean	Dev	3	6,449	2,149.7	1,821.0
Korean	Test	3	7,142	2,380.7	1,710.0
Korean	Train	24	76,440	3,185.0	2,065.0
Korean	All	30	90,031	3,001.0	1,765.5
Spanish	Dev	4	6,750	1,687.5	1,325.0
Spanish	Test	4	7,789	1,947.2	1,593.5
Spanish	Train	36	110,547	3,070.8	1,408.0
Spanish	All	44	125,086	2,842.9	1,408.0

Table 5: Token statistics per language and split.

C Further Dataset Statistics

Table 5 reports token-level statistics for each language in the corpus, broken down by data split (dev, test, train) and aggregated across all splits. For each partition, we report the number of stories, the total token count, and the per-story mean and median lengths. The aggregate row summarises statistics across all splits for a given language.

D Data Consent

During the corpus acquisition process, particularly for extending the Korean and Dutch fanfiction collections that are less well represented in AO3, we actively sought input from fan authors and readers to understand their concerns and establish best practices for data use. We engaged in open dialogue on licensing, consent, and data-sharing options. Acknowledging the fan community's concern towards commercial usage of their works, we publicly share only the full content of stories without copyright or copyright works with the author's explicit knowledge and consent that their works will be shared publicly as open data for research. To get enough

```
11_do_not_wake_love.txt 0   0   “       -   -   -   -   -   -   -   -
11_do_not_wake_love.txt 0   1   C       -   -   -   -   -   -   -   -
11_do_not_wake_love.txt 0   2   ’       -   -   -   -   -   -   -   -
11_do_not_wake_love.txt 0   3   mon     -   -   -   -   -   -   -   -
11_do_not_wake_love.txt 0   4   <EMPTY> -   -   -   -   -   -   -   (0)
11_do_not_wake_love.txt 0   5   .       -   -   -   -   -   -   -   -
```

Figure 1: Example sentence in CoNLL 2012 format. Note the dummy token <EMPTY>, which refers to the addressee.

```
# sent_id = 11-39
# text = “C’mon.
1   “   ''  PUNCT   ``  _           2   punct   2:punct _
2   C   C   PROPN   NNP Number=Sing 0   root    0:root  _
3   ’   '   PUNCT   ''  _           2   punct   2:punct _
4   mon mon X       FW  _           2   flat    2:flat  _
4.1 _   _   _       _   _           _   _       4:null  Entity=(e0)
5   .   .   PUNCT   .   _           2   punct   2:punct _

# sent_id = 11-40
# text = We’ll take my car to go deal with this ghost.
1   We    we    PRON  PRP  Case=[...]   4   nsubj     4:nsubj
                                                      Entity=(e4)|SplitAnte=e3<e4,e0<e4
2   ’     's    AUX   MD   VerbForm=Fin 4   aux       4:aux   _
3   ll    will  AUX   MD   VerbForm=Fin 4   aux       4:aux   _
4   take  take  VERB  VB   VerbForm=Inf 0   root      0:root  _
5   my    my    PRON  PRP$ Case=[...]   6   nmod:poss 6:nmod:poss Entity=(e3)
6   car   car   NOUN  NN   Number=Sing  4   obj       4:obj   _
7   to    to    PART  TO   _            8   mark      8:mark  _
8   go    go    VERB  VB   VerbForm=Inf 4   advcl     4:advcl _
9   deal  deal  VERB  VB   VerbForm=Inf 8   xcomp     8:xcomp _
10  with  with  ADP   IN   _            12  case      12:case _
11  this  this  DET   DT   Number=[...] 12  det       12:det  _
12  ghost ghost NOUN  NN   Number=Sing  9   obl       9:obl   _
13  .     .     PUNCT .    _            4   punct     4:punct _
```

Figure 2: Example sentences in CorefUD format.

consent, we contacted more than 300 fanfiction authors in Dutch and Korean. The participant sheet and consent form that we used to contact authors can be found in the repository. The data collection protocol was approved by the ethics review board of the University of Groningen.