Towards universal annotation guidelines for coreference resolution and information status

Andrew Dyer
Department of Language Science and Technology
Saarland University
andrew.dyer@uni-saarland.de

Overview

As a computational linguist studying word order typology using quantitative methods, **I want to** develop a set of guidelines for annotating coreference resolution and information status in texts in all the world's languages. It should be built natively in Universal Dependencies, as a layer on top of UD annotation, **so that** I and other researchers can use it to study the interaction between discourse and word order in the world's languages.

These guidelines will integrate into CorefUD (Nedoluzhko et al., 2022) and Universal Anaphora (Poesio et al., 2024), so that they will be compatible with existing annotation and evaluation tools.

The guidelines will be used to annotate on top of UD treebanks in any language. They will thus be *universal* and *cross-lingually consistent*. The resulting collection will be a collaborative, open-source and open-contribution resource, such that contributors can add to it and deliberate on changes to the annotation guidelines in a similar manner to Universal Dependencies (Nivre et al., 2020).

This project will benefit NLP, computational linguistics, and digital humanities researchers by providing an open resource for large scale cross-linguistic study, including languages that are low-resource or receive less attention in NLP.

Background

Our interest in this resource stems from a desire to study the interaction of coreference, information status, and syntax. Experimental research has shown how information status – broadly, whether entities in the discourse are *given* or *new* – licenses non-canonical structures in many languages (Skopeteas et al., 2010), and some languages are known to be *discourse configurational* (Kiss, 1995) in the sense that word order choice is determined by information status. Computational typological study is enabled by some existing corpora such as the Prague Discourse Treebank (Synková et al., 2022) and GUM (Zeldes, 2017). However, there are currently no cross-lingually consistent multilingual corpora to enable study of such discourse-syntactic phenomena at a large scale.

Coreference resolution has become an increasingly multilingual task, with shared tasks focusing on models that can parse a variety of languages (Žabokrtský et al., 2023). Along with this, there have been initial efforts at devising unifying schemes of annotation, such as *Universal Anaphora.*[1] A goal of Universal Anaphora 2.0 (Poesio et al., 2024) is *"to start attempting developing common guidelines, as done in UD, ideally in collaboration with the linguistic community."*

Our aim here is to facilitate such a gathering of linguists to devise a set of universally applicable guidelines that would suit the need of study in all languages. The current project would build on Universal Anaphora in such a way that information status is included as an annotation feature.

[1] https://universalanaphora.github.io/UniversalAnaphora/

Open questions related to the project

(Poesio et al., 2024) have listed several phenomena that are supported by the Universal Anaphora format and scorer, including split antecedents and bridging, which we need to model in a cross-linguistically consistent manner. Of particular interest to us is the treatment of zero anaphora[2], which are included in some CorefUD corpora as zero tokens. As pro-drop is a core feature of many languages and heavily influenced by information status and dialogue flow, we consider it important to model this in all languages, but it is tricky to maintain a set of rules for inserting such dropped arguments that applies parsimoniously to all languages, whether or not, for example, they index arguments through verbal or nominal morphology (Haspelmath, 2013).

Our work on this project so far has used modern prose, which includes a lot of dialogue and speaker perspectives. Most existing datasets assume the *omniscient reader* perspective, with coreference and information status assigned based on the reader's knowledge[3]. We wonder whether a speaker-centred perspective would more truly reflect information status in language production.

How participation in the Training School will be useful for the project

The hackathon will bring together UD-interested linguists from a wide variety of linguistic backgrounds, including those that typically receive less attention in NLP. This is essential for being able to devise annotation guidelines that meet the objective of universality, as without their input guidelines would be prone to bias towards dominant languages.

As a project manager of this annotation effort, I will also benefit from greater understanding of the technical infrastructure and asset management supporting Universal Dependencies and how it applies to this project. While I have worked with Git professionally and have managed repositories with documentation, the requirements of Universal Dependencies are quite particular and complex.

Current Status

We have begun annotation of a parallel parsed corpus of modern prose for the purpose of computational typological study. This corpus contains more than 40 languages, of which six now have some annotation: English (2.3k sentences); Greek (830 sentences); Indonesian (170 sentences); Portuguese (440 sentences); Turkish (150 sentences); and Ukrainian (350 sentences).

We have annotation guidelines relevant to our project, which we have tried to craft to be consistent and comparable between languages. These guidelines are continually evolving as we add more languages and texts. However, they fall short of our vision in a few ways:

1. While the corpus is UD-parsed, annotators have not had access to this layer of annotation when annotating. This means that there may be incongruencies between syntactic and coreference annotation that we would seek to avoid.
2. There are elements of CorefUD annotation that we have either avoided or approached in a more reduced form (e.g. bridging).
3. While we have applied our linguistic knowledge for the languages that we have worked on, we cannot guarantee that our guidelines will meet the needs of all languages.

We look forward to sharing our progress so far and receiving feedback, so that our guidelines can develop from a local project into a truly universal resource.

[2] See Feature GB522 of Grambank (Skirgård et al., 2023).

[3] For example, the *FantasyCoref* corpus (Han et al., 2021) makes this decision, while the authors note that different states of knowledge between characters is a staple of fictional prose.

References

Han, S., Seo, S., Kang, M., Kim, J., Choi, N., Song, M., & Choi, J. D. (2021). FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View. In M. Ogrodniczuk, S. Pradhan, M. Poesio, Y. Grishina, & V. Ng (Eds.), *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference* (pp. 24–35). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.crac-1.3

Haspelmath, M. (2013). Argument indexing: A conceptual framework for the syntactic status of bound person forms. In *Argument indexing: A conceptual framework for the syntactic status of bound person forms* (pp. 197–226). De Gruyter Mouton. https://doi.org/10.1515/9783110331127.197

Kiss, K. É. (Ed.). (1995). *Discourse configurational languages*. Oxford University Press.

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., Zeldes, A., & Zeman, D. (2022). CorefUD 1.0: Coreference Meets Universal Dependencies. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4859–4872). European Language Resources Association. https://aclanthology.org/2022.lrec-1.520

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4034–4043). European Language Resources Association. https://aclanthology.org/2020.lrec-1.497

Poesio, M., Ogrodniczuk, M., Ng, V., Pradhan, S., Yu, J., Moosavi, N. S., Paun, S., Zeldes, A., Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., & Zeman, D. (2024). *Universal*

Anaphora: The First Three Years. https://ufal.mff.cuni.cz/biblio/attachments/2024-nedoluzhko-p5075542261083397686.pdf

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., … Gray, R. D. (2023). Grambank Reveals Global Patterns in the Structural Diversity of the World's Languages. *Science Advances*, *9*(16). https://doi.org/10.1126/sciadv.adg6175

Skopeteas, S., Fanselow, G., Asatiani, R., Bartaia, S., Khizanishvili, T., & Kvaskhvadze, T. (2010). *Effects of givenness and constraints on free word order*.

Synková, P., Rysová, M., Mírovský, J., Poláková, L., Sheller, V., Zdeňková, J., Zikánová, Š., & Hajičová, E. (2022). Prague Discourse Treebank 3.0. *https://ufal.mff.cuni.cz/pdit3.0*. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4875

Žabokrtský, Z., Konopik, M., Nedoluzhko, A., Novák, M., Ogrodniczuk, M., Popel, M., Prazak, O., Sido, J., & Zeman, D. (2023). Findings of the Second Shared Task on Multilingual Coreference Resolution. In Z. Žabokrtský & M. Ogrodniczuk (Eds.), *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution* (pp. 1–18). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.crac-sharedtask.1

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612. https://doi.org/10.1007/s10579-016-9343-x