A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari,
Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska,
Syahidah Asma Umniyati, Helena Rodrigues Menezes de Oliveira Vaz

Language Science and Technology
Saarland University
Saarbrücken, Germany
{firstname.lastname}@uni-saarland.de

Abstract

Information status — the newness or givenness of referents in discourse — is known to affect the production of language at many different levels. At the morphosyntactic level, information status gives rise to special words orders, elisions, and other phenomena that challenge the notion that morphosyntax can be considered independent of discourse context. Though there are many language-specific corpora annotated for information status and its related phenomena, coreference and anaphora resolution, what is not available at present is a cross-lingually consistently annotated corpus or annotation scheme that would allow for comparative study of these phenomena across many diverse languages. In this paper we present our work to build such a resource. We are annotating a parsed, parallel corpus of prose in many languages for information status and coreference resolution, so that like-for-like cross-lingual comparisons can be made at the intersection of discourse and syntax. Our corpus can and will be used both for corpus analysis and for model training.

1 Introduction

When speakers[1] produce sentences, utterances and meanings, they usually do so not in isolation, but in the context of a longer discourse and in a communicative context between speakers with shared knowledge of the world that coincides or differs in important ways. The shared world knowledge between speakers mediates what meaning can be interpreted from utterances (Beyer, 2015), while common ground in conversation mediates what information need be explicitly stated (Karttunen, 1974).

Central within this dynamic is information status: broadly, whether information communicated is *new* – that is, being encountered or asserted for the first time; or *given* – the information has been introduced before, or is otherwise already inferrable by the receiver (Chafe, 1976). Broadly speaking, information, referents and arguments that are considered known in the common ground of the discourse may be reordered, reduced, receive special (intonational) markers, or may even be omitted altogether in the aid of information flow, allowing processing time and emphasis for the assertion of more novel or surprising information (Fenk-Oczlon, 2001).

How this plays out varies widely across languages. Languages such as English have definiteness as a grammatical feature within the noun phrase, thus allowing their hierarchy of givenness to be visible through word forms (Gundel et al., 1993). Other languages, such as Czech and Hungarian, convey the givenness of information though word order, and are considered *discourse configurational* languages as a result of this expectation (Kiss, 1995). Additionally, many languages allow for given information to be omitted from the sentence entirely, either relying on indexing arguments through morphological processes on root words, or by relying on speakers to infer arguments from context. Japanese is an example of such a language (Vermeulen, 2012). In these languages, information flow is handled by simple elision of overt arguments.

The role of information status on language production has been well studied in individual languages using both psycholinguistic experimentation and corpus study. Seminal studies include Arnold et al. (2000) on word order in English, Skopeteas and Fanselow (2010) on cross-linguistic differences in the expression of focus, and Wang et al. (2012) on the so-called Chomsky illusion, showing how focus is a determinant of depth of syntactic processing in Mandarin Chinese.

There has also been increasing interest in cross-lingual comparison of the way information status

[1] We use *speakers* as it is the term for those who produce language that will be most readily understood, but these arguments apply equally to signed languages, as well as the written modality.

Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024), pages 55 - 64
December 5-6, 2024 ©2024 Association for Computational Linguistics

is signalled and the way it affects language production in the world's languages.

To study the influence of information status on syntax cross-lingually – for example, the shifting of given information to a sentence-initial position, or the use of pronominal forms for an entity that is currently active in the discourse – we need corpora that are multilingual and consistently annotated both for syntax and information status (Lüdeling et al., 2014).

Information status is closely related to the task of coreference and anaphora resolution: the identification of expressions in a text that refer to the same entity, and there are several corpora that combine these two tasks (Markert et al., 2012; Zeldes, 2017). In the interests of cross-lingual natural language processing, there have been efforts to bring diverse corpora for coreference and anaphora resolution together into a common format (Nedoluzhko et al., 2022), and there are beginning efforts towards consistent multilingual annotation (Poesio et al., 2024). However, as of yet there is no resource that fully meets the criteria that we need met in order to pursue multilingual comparative studies.

We introduce our work to develop such a resource. We annotate on top of a parallel corpus of modern literature in translation, predictively parsed according to Universal Dependencies annotation. We annotate spans of entity mentions, with coreference chain annotation to track mentions of the same underlying entity; and information status and mention type annotations to describe the mention. In this way, we can use the underlying syntactic annotation of sentences to follow the placement of referring expressions, to quantify how the information status of such expressions, their mention type, and the recency of mentions of the same entity in the discourse, affect the order in which they are placed.

We annotate texts in a diverse variety of languages, with common annotation guidelines applying to each language that is added. As each new language is added, we work to ensure that our annotation principles and guidelines apply consistently to each language, ensuring that like-for-like comparisons can be made betweeen languages.

Our corpus has the following benefits:

- **Parallel**: The texts used in the corpus are direct translations of works of prose in each language. This makes it easier to make direct comparisons of phenomena between languages.
- **Minimalist**: We are conservative with regard to mention spans, including only the most relevant information and minimising overlap. This makes annotation easier and faster, and visually clearer for users and programs.
- **Feature modularity**: We make features modular, increasing the efficiency and precision of annotation. This allows flexible and granular descriptions of mentions while avoiding feature explosion, and is simpler for annotators and readers than a lengthy list of features.

In this paper, we will describe and motivate our annotation scheme in the context of existing resources; and discuss our current workflow and progress in annotation.

2 Related Work

There are many monolingual corpora for coreference resolution and/or information status that have been used for quantitative study of the effects of word order. For example, *RefLex* (Baumann and Riester, 2012) and *ISNotes* (Markert et al., 2012) are corpora in German and English respectively, with span annotation of entity mentions, coreference links, and nuanced categories of mention type. OntoNotes (Hovy et al., 2006) is among the most widely used corpora for coreference and anaphora resolution, and covers English, Arabic and Mandarin Chinese. The *Georgetown University Multilayer* corpus (*GUM*) (Zeldes, 2017) is a multimodal corpus of English annotated with UD syntactic struture, coreference, and information status, among many other layers of annotation. The information status and mention type labels of *GUM* are inherited by our scheme.

Many coreference resolution corpora — including *GUM* – from a variety of European languages have been assembled and harmonised in *CorefUD* (Nedoluzhko et al., 2022), where coreference annotation is joined with predictive Universal Dependencies parsing. The harmonisation of many schemes into a common format has been the basis of considerably many experiments and advances in training cross-lingual and multilingual coreference resolution models (Ogrodniczuk et al., 2023).

Universal Anaphora[2] (Poesio et al., 2024) is a Universal Dependencies-inspired effort to create a common framework for annotation of coreference resolution so that coreference and information sta-

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

tus can be compared across languages in a similar manner to Universal Dependencies. As of the time of writing, Universal Anaphora has contributed an enhanced file format for representation of coreference resolution (the conll-UA format), and a wide range of tools for scoring and validation of coreference resolution models, but work to create a common linguistic annotation scheme has not yet been undertaken.

To our knowledge, there are no currently existing *parallel* multilingual corpora annotated for both coreference resolution and information status, and this is where we seek to make our contribution.

3 Data and format

3.1 Data

The corpus that we use as the base for our annotation is *mini-CIEP+* (Verkerk and Talamo, 2024). *mini-CIEP+* is a multilingual parallel[3] corpus of modern prose in translation. The corpus is predictively parsed according to Universal Dependencies (Nivre et al., 2020) using Stanza (Qi et al., 2020).[4] The corpus is thus represented in conllu format[5]. The corpus covers 40 languages at the time of writing, with more to be added.

From among this data, we have annotated data from books in seven languages: English, Ukrainian, Modern Greek, Portuguese, Hindi, Turkish, and Indonesian. The choice of these languages is motivated by linguistic diversity: the languages come from a variety of families (Indo-European, Turkic, Austronesian), and exhibit varying degrees of word order freedom, morphological indexing and pro-drop. Accommodating these languages early on allows us to address the linguistic challenges that arise from them.

The data being drawn from the literary domain presents its own challenges. Compared with the more formal styles favoured in many resources such as OntoNotes and GUM, the literary genre includes complicated annotation issues such as asymmetry of knowledge between characters, changes in entities, and lexical variation in entity description (Han et al., 2021). The benefit of this challenge is that we expect to encounter more idiosyncratic and diverse language use, which has benefits both for diversity of sampling and for model training.

3.2 Format

Our annotation of the corpus is output in the *CorefUD* format[6]. The CorefUD format follows that of conllu, but places mention span annotation in the *misc* column, along with other data concerning coreference relations. The building blocks in this format are *spans* and *clusters*. Mentions of entities in the discourse are represented by a tuple-like span object, opening on the token where the span begins, and closing on the token where it ends. Within this span is contained an entity ID, specifying the ID of the underlying entity of the mention, as well as various other attributes. We refer to the CorefUD file format description for more details and examples, but we give an example of the output of our corpus in Fig 1.

We choose to follow this format as closely as possible so as to be able to integrate our corpus with existing resources, including CorefUD corpora and evaluation scripts, so that we can train models to parse more of the corpus and further corpora.

4 Annotation design

4.1 General principles

In the interests of speedy annotation, and to avoid overburdening annotators with too many labels, we try to keep our labels simple and modular. That is to say, that rather than giving annotators a deep hierarchy of labels to choose from, we aim to give a set of attributes with limited options, as shown in Table 1.

For example, we only use two labels for information status: *given* and *new*. There are finer grained measures of coreference, such as the near-identity relations used by Recasens et al. (2011), that we do not include. We also do not include focus, often cited as a central part of information structure, due to the difficulty of defining this in a cross-lingually satisfactory way (Matić and Wedgwood, 2013).

4.2 Markables

4.2.1 Markable spans

A markable is a span of text that may constitute an entity mention. (Dipper et al., 2007) The following structures are always annotated as markables:

[3] Parallel in the sense that the same work is represented - either in original or translation - in each language, and thus the context is the same. The texts are not strictly bitexts, or sentence- or token-aligned, but contain in theory the same content, ensuring comparability.

[4] https://stanfordnlp.github.io/stanza/

[5] https://universaldependencies.org/format.html

[6] https://ufal.mff.cuni.cz/~popel/corefud-1.0/corefud-1.0-format.pdf

```
# sent_id = Alquimista_English_006_2
# text = During the two hours that they talked, she told him she was the merchant's daughter ...
1       During during ADP    IN     _       4       case    _       TokenRange=74:80
2       the     the    DET    DT     Definite=Def|PronType=Art    4       det     _       Entity=(e7-time-3-CorefType:coref,InfStat:new|TokenRange=81:84
3       two     two    NUM    CD     NumForm=Word|NumType=Card    4       nummod _       TokenRange=85:88
4       hours   hour   NOUN   NNS    Number=Plur    10      obl     _       Entity=e7)|TokenRange=89:94
5       that    that   PRON   WDT    PronType=Rel   7       obj     _       TokenRange=95:99
6       they    they   PRON   PRP    Case=Nom|Number=Plur|Person=3|PronType=Prs 7       nsubj  _       Entity=(e8-person-1-CorefType:ana,InfStat:given)|
       SplitAnte=e2<e8,e1<e2|TokenRange=100:104
7       talked talk   VERB   VBD    Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin 4   acl:relcl      _       SpaceAfter=No|TokenRange=105:111
8       ,       ,      PUNCT  ,      _       10      punct  _       TokenRange=111:112
9       she     she    PRON   PRP    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs 10  nsubj  _       Entity=(e2-person-1-CorefType:ana,InfStat:
       given)|TokenRange=113:116
10      told    tell   VERB   VBD    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 0   root   _       TokenRange=117:121
11      him     he     PRON   PRP    Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs 10 iobj   _       Entity=(e1-person-1-CorefType:ana,InfStat:
       given)|TokenRange=122:125
12      she     she    PRON   PRP    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs 17  nsubj  _       Entity=(e2-person-1-CorefType:ana,InfStat:
       given)|TokenRange=126:129
13      was     be     AUX    VBD    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin 17  cop    _       TokenRange=130:133
14      the     the    DET    DT     Definite=Def|PronType=Art    15      det     _       Entity=(e2-person-4-CorefType:pred,InfStat:new(e3-person-2-
       CorefType:coref,InfStat:given|TokenRange=134:137
15-16  merchant's    _      _      _      _      _       _       _       Entity=e3)|TokenRange=138:148
15      merchant       merchant       NOUN   NN     Number=Sing    17      nmod:poss      _       _
16      's      's     PART   POS    _       15      case   _       _
17      daughter       daughter       NOUN   NN     Number=Sing    10      ccomp  _       Entity=e2)|SpaceAfter=No|TokenRange=149:157
```

Figure 1: An example sentence from our corpus (from the English portion) in CorefUD format, with entity annotation in the *misc* column. Note that mention spans may open on one token and close on another, and that two mentions may start or end on the same token (but may not cross eachother).

- Referring pronouns (excluding dummy pronouns and relative pronouns)
- Referring noun phrases (excluding idiomatic instances)

Additionally, we annotate as markables these structures if they are coreferred by an anaphoric mention:

- Interrogative and quantifying pronouns, e.g. *whoever*, *anything*
- Verbal and other non-nominal phrases that are referred to anaphorically as discourse deixis (Dipper and Zinsmeister, 2009); for example *"[He said no]. **[That]** surprised me"*
- Pro-adverbs such as *here* and *then*

Pronominal clitics may also be annotated as markables provided that they are not part of an introverted reflexive verb phrase (Haspelmath, 2008). These are common in many Indo-European languages such as Portuguese and Dutch, where they simply reinforce that the agent of a verb is the same as its patient.

The greatest divergence with most other schemes in *CorefUD* in terms of annotation philosophy is that we are more minimalist with what we include in a markable. Such schemes typically cover the full syntactic noun-phrase, including all determiners, modifiers, adjuncts and clausal expansions. By contrast, we opt for an approach where only the most relevant information used to identify the entity is included. This always includes the syntactic head, but adjuncts and modifiers are only included if they provide information that is essential to understanding and referring to the referent.

We use some linguistic tests to decide on what information should be included when annotating a markable span:

- *Question test*: If we form a question to which the entity being referred to is the answer, would the same wording typically be used in the answer?
- *Repeated mention test*: Would the same wording be used (or is it used) in a subsequent mention to refer to the entity?
- *Contrast test*: Does the wording of the mention serve to contrast this referred entity with another similar entity?

Likewise, we also do not include possessive pronouns as part of the markable span (but may include them in their individual spans, see ex (1))[7], and we do not mark conjunctions as a single markable (see ex (2)). We are just as often interested in the order of possessor and possessum in such expressions, and if we need the full expression, it is easy to recover this from the dependency tree.

(1) a. [Our] [house] is on fire
b. ∗ [[Our] house] is on fire

(2) a. [Tom], [Dick] and [Harry] were there.
b. ∗ [[Tom], [Dick] and [Harry]] were there.

4.2.2 Zero anaphora

Many corpora in CorefUD use *zero tokens* to represent dropped or omitted arguments of verbs, or

[7] In this paper we use ∗ to indicate that we do not identify mentions this way; not that the text itself is ungrammatical.

in some case of nouns. For example, the Ancora corpus for Spanish (Taulé et al., 2008) annotates the referent of indexed subjects of inflected verbs; the SzegedKoref corpus for Hungarian annotates indexed subjects and objects of verbs and possessors of nouns (Vincze et al., 2018); and the Turkish ITCC corpus also annotates indexed subjects and possessors, potentially leading to multiple mention spans to be annotated on the same token (Pamay Arslan and Eryiğit, 2025).

The languages to which this is applied are typically those with extensive pro-drop, and particularly those where arguments and possessors are indexed with morphology on the verb or noun. The motivation behind this is that in such languages, indexed arguments constitute the majority of anaphoric expressions. In terms of information status, the topicalisation of arguments is also a factor in whether an overt pronoun is used or not (Givón, 1983).

To make an annotation scheme *universal*, we believe that this needs to be accommodated for all languages. Our corpus includes, on one end of the spectrum, languages such as English, which allows only minimal and restricted pro-drop; and on the other end, languages such as Turkish and Portuguese, which employ extensive and free prodrop. In both cases, we allow for the annotation of dropped arguments as zero tokens with the following conditions:

1. The expression's syntactic role supports a category relevant to the argument.
2. The argument is indexed through morphology on the expression, however minimally.
3. The argument is not overtly mentioned in the same or a head clause.

Keeping to these rules allows us to apply zero tokens to any language while maintaining like-for-like comparisons, and is less burdensome for annotators.

4.3 Coreference relations

The basic coreference relation in our corpus is identity. This is a symmetric relation that implies that the entity referred to by mention A is one and the same as the one that is referred to by mention B. In the output, identity coreference is represented by two mentions sharing the same entity ID: in other words, a cluster representation. For two entities to be identity coreferential, they must share the same underlying entity. An anaphoric mention, for example, will have the same entity ID as its antecedent. This may also apply to both mentions in a predicative statement. For example, in ex (3), all three mentions are identity coreferent. Likewise, two appositional mentions may also be identity coreferent, as in ex (4)

(3) This is John Snow$_1$, he$_1$'s King in the North$_1$.

(4) Narcissus$_1$, a youth$_1$ who knelt daily...

Split antecedence is also represented in our annotation scheme. Unlike identity coreference, this is an asymmetric relation that signifies that the entity referred to by mention A is a superset to one or more antecedent entities. For example These entities may be in conjunction or free configuration (Yu et al., 2020). In the output, split antecedence is represented using the SplitAnte feature in the CorefUD format. An example of this can be seen in Fig 1.

4.4 Attributes

Key attributes relating to information status, mention type and other important linguistic phenomena are carried in attributes annotated onto mention spans. These are represented in key-value pairs, and these are listed in Table 1.

Our motivating principle for the attributes is modularity. While each of the attributes is quite simple, reducing the effort at annotation time, combinations of attributes may build a granular description of the mention's characteristics, while avoiding combinatorial explosions of discrete features. Modularity also allows flexibility in annotation, giving greater freedom to annotate unusual mentions.

Attribute	**Values**	**Required**
InfStat	new, given	true
CorefType	ana, cata, pred, disc, appos, coref	true
Indexing	NullSubj, NullObj, NullPoss	false
Bridging	*boolean*	false
Deixis	*boolean*	false

Table 1: The set of attributes that we can apply to a mention.

4.4.1 Information Status

We use only two values for information status: *new* and *given*. Unlike some other schemes (e.g. GUM), we do not include *accessible* – i.e. a mention of an entity that is considered given simply due to cultural or environmental context, such as *God* or *the sky* – as a tag. The reason for this is that it is difficult to fully define cross-lingually what can be considered accessible, due to the different cultural contexts of each book.

We apply information status to *mentions*, not to entities themselves. The *new* value applies to the first mention of an entity, but it may also apply to another mention that substantially expands the known information about that entity. In ex (5), that the referent is named Jon Snow and that he is King in the North is *new* information, even if the entity is already introduced. We consider this more reflective of human packaging and processing of information, recognising that a speaker might employ information status-related strategies to convey this new information.

(5) [This](new,cata) is [Jon Snow](new,pred). [He](given,ana)'s [King in the North](new,pred).

4.4.2 Coreference type

CorefType is the attribute that we use to classify the type of mention: for example, an anaphoric mention such as a pronoun; a predicate mention, such as in an *is* statement (e.g. ex (3)); or a general *coref* mention, for all kinds of open class referring expressions. We inherit the coreference types used in *GUM* for mentions which are coreferent with another mention. These are *ana* (anaphor); *cata* (cataphor); *pred* (predicate); appos (apposition); *disc* (discourse deixis); and *coref* (lexical coreference). These are applied to individual mentions, and may also be applied to singletons (sole mentions of an entity).

4.4.3 Deixis

To investigate the effect of deixis in conjunction with givenness, we introduce the attribute *Deixis*. This applies to any deictic mention of an entity; one where the reference can only be fully understood from the spatiotemporal perspective of the utterer. This attribute applies to:

- Any anaphoric first- or second-person reference. These are also considered *given* from the utterer's perspective, giving all such mentions the combination *(given, ana, deixis)*.
- Spatial demonstratives, such as *here, over there*, and nouns with spatial determiners such as *that guy, this place* if relying on the location of the utterer. Such references may be either new or given, depending on the context.
- Temporal adverbs or noun phrases, such as *now, yesterday, next year*.

4.4.4 Bridging

Bridging refers to a relation between two entities in a discourse where a target entity is not strictly the same as its antecedent, but bears a strong semantic link and is inferrable (Clark, 1977). In ex (6), *the trees* is inferrable as part of *the woods*, and therefore does not need introducing in the same way that other entities might.

(6) Lost in **the woods**, **the trees** devour me.

In *CorefUD* corpora that include it, bridging is a relation between two entities, requiring a link from one mention to another. It is represented, like split antecedence, by a pointer from the entity ID of the mention to the entity ID of the antecedent.

Though we are interested in bridging, since it affects the manner in which entity mentions are introduced, for the sake of simplicity we represent bridging as a boolean attribute which applies only to the target mention. The antecedent, from which the target is inferrable, is not annotated. This choice is motivated by the need for rapid annotation and simplicity among a mixed team of annotators. Finding the antecedent of a bridging mention is often difficult, and indeed the antecedent may not be a nominal at all, but may only appear at a phrasal or even discourse level. The representation of bridging is thus contained within the mention span, rather than in the *misc* column. Table 2 shows an example of our bridging annotation.

Lost	
in	
the	Entity=(e1-object-2-InfStat:new
woods	Entity=e1)
,	
the	Entity=(e2-object-2-InfStat:new,Bridging:True
trees	Entity=e2)
devour	
me	

Table 2: An example of bridging annotation in our current scheme. Bridging is represented as a boolean value, without pointing to the antecedent; and is annotated within the mention span, rather than in the misc column.

4.4.5 Indexing

As explained in Section 4.2.2, in many languages anaphoric subject, agent, patient or nominal possessor arguments are indexed through morphology on the syntactic head phrase, with the option of omitting a (pro)nominal mention.[8]

These indexed mentions are included as zero tokens, and we use the attribute *Indexing* to identify the argument that they index. The three basic types are inherited from the CorefUD scheme:

1. `NullSubj`: An indexed subject
2. `NullObj`: An indexed object
3. `NullPoss`: An indexed possessor

5 Annotation Procedure

We performed our annotation using Brat (Stenetorp et al., 2012)[9], hosted on a webserver. Brat was chosen primarily for its ease of use and customisation of the configuration.

The annotators are each native speakers of the language that they annotate. All annotators begin by annotating sentences in English to practice, with a native English-speaker reviewing, before moving on to their own languages. Practice annotation in English is done collaboratively and different annotators' decisions are compared. Once annotators are confident of their understanding of the guidelines, they move on to annotation in their own languages. Again, we keep an open forum for discussion of linguistic issues that arise in new languages, and policies evolve based on new linguistic scenarios encountered.

Annotating the full text of each book in one document would be impossible due to the limitations of Brat (and other coreference annotation software): the sheer amount of text and arrows between elements would overwhelm the GUI. For this reason, we chunk each book in each language into chunks of 10 sentences each, and perform annotation on each chunk. Information status is carried over between chunks, so that an entity that has been seen in a previous chunk of the same book will be considered *given* in its next appearance. Attachment of coreference chains between chunks, however, is a task that will need to be completed later.

6 Progress

At the time of writing, our progress in number of sentences annotated is as shown in Table 3. We have completed scripts to serialise from Brat annotation to CorefUD format, and so can output this data in the appropriate format to be used in CorefUD scripts.

Language	**Sentences annotated (approx.)**
English	3130
Portuguese	2320
Greek	900
Ukrainian	750
Indonesian	190
Hindi	270
Turkish	130

Table 3: Approximate number of sentences annotated per language covered so far, as of November 2024.

7 Future Plans

Now that we have a large amount of annotated data in several languages, we are closer to being able to train multilingual models to predictively annotate data and speed up our annotation process (Pražák and Konopik, 2022), as well as to evaluate the consistency and intrinsic strengths of our annotation (Chai and Strube, 2023). We may also apply techniques such as annotation projection to speed up pre-annotation.

A major shortcoming of our work so far is that we have not instituted quantitative quality control measures such as inter-annotator agreement. One reason for this is that we have only one annotator for each language other than English; while for English we prioritised annotating as much data as possible. Another is that the annotation platform, Brat, does not easily facilitate such measures or annotation of the same data by multiple annotators. In the long term we would like to move to another tool such as INCEpTION[10] (Klie et al., 2018), which would facilitate this when we are able to recruit more annotators.

A shortcoming of Brat is that annotators are unable to see the underlying syntax trees when annotating. Since our goal is to be able to analyse syntax and discourse annotation together, it would be beneficial to ensure that, for example, annotation spans do not cross subtree boundaries (Popel et al., 2021).

Finally, the shortness of our chunks is a problem for studying long range coreferences, and an important next step is to concatenate chunks to form full

[8] See features GB089-GB094 in Grambank (Skirgård et al., 2023) for descriptions and a list of languages with these features.

[9] `https://brat.nlplab.org/`

[10] `https://inception-project.github.io/`

documents for single works and to link coreference chains referring to the same entities.

8 Conclusion

We have presented our annotation scheme, design, and ongoing work on a multilingual corpus that will enable large scale corpus-based analyses of the interplay of information status and word order in a cross-section of the world's languages. Our corpus is now at the stage where we can experiment with model training and evaluation, with sentences annotated in seven languages so far, and annotation guidelines continuously evolving to meet the demands of new languages. We look forward to our first release and to the first applications of the data to answer questions regarding the intersection of information status, information theory, and word order variability.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

Special thanks to Martin Popel, who provided an improved script to convert the Brat annotation to a form compliant with CorefUD.

Ethics

Data availability

Our data contains annotations of works which are protected under copyright. As a result of this we cannot make our corpus open-source and open-access. However, the copyright law of our country allows us to share portions of copyrighted works with researchers for non-commercial purposes, and we are happy to do this on request per the conditions explained in Verkerk and Talamo (2024).

Annotators

Annotation was performed variously by members of the research group working on this project, collaborators in other departments, and currently enrolled students at our institution who were employed under a student assistant contract.

Student assistants were recruited via a call for assistants circulated by email within our institution. Shortlisted candidates were interviewed, and from these candidates annotators were selected based on the interview, their linguistic experience, and their language skills. Student assistant annotators were paid above the minimum wage of our country, and working time was flexible and limited to be compatible with the demands of full-time study.

All annotators, regardless of status, played an important role in the project and were treated with respect and kindness. All were offered to be named as co-authors and are so named in this paper.

References

Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects.

Christian Beyer. 2015. Meaning, Context, and Background. In Thomas Metzinger and Jennifer M. Windt, editors, *Open MIND, 2-vol. set*. The MIT Press.

Wallace L Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*.

Haixia Chai and Michael Strube. 2023. Investigating multilingual coreference resolution by universal annotations. *arXiv preprint arXiv:2310.17734*.

Herbert H Clark. 1977. Bridging. In *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

Stefanie Dipper, Michael Goetze, and Stavros Skopeteas, editors. 2007. *Information structure in cross-linguistic corpora : annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Universitätsverlag Potsdam.

Stefanie Dipper and Heike Zinsmeister. 2009. Annotating discourse anaphora. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 166–169, Suntec, Singapore. Association for Computational Linguistics.

Gertraud Fenk-Oczlon. 2001. Familiarity, information flow, and linguistic form. In *Frequency and the Emergence of Linguistic Structure*. John Benjamins.

Talmy Givón. 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study*. John Benjamins, Amsterdam; Philadelphia. 2010.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View. In *Proceedings of the Fourth Workshop on*

Computational Models of Reference, Anaphora and Coreference, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Haspelmath. 2008. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery, v.6, 40-63 (2008)*, 6.

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.

Lauri Karttunen. 1974. Presupposition and linguistic content. *Theoretical Linguistics*, 1(1-3):181–194. Publisher: De Gruyter Mouton Section: Theoretical Linguistics.

Katalin É Kiss, editor. 1995. *Discourse configurational languages*. Oxford studies in comparative sytax. Oxford University Press, New York.

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. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Anke Lüdeling, Julia Ritz, Manfred Stede, and Amir Zeldes. 2014. *Corpus linguistics and information structure research*. Oxford, Oxford University Press.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Dejan Matić and Daniel Wedgwood. 2013. The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis. *Journal of Linguistics*, 49(1):127–163.

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, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, and Massimo Poesio, editors. 2023. *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*. Association for Computational Linguistics, Singapore.

Tuğba Pamay Arslan and Gülşen Eryiğit. 2025. Enhancing Turkish Coreference Resolution: Insights from deep learning, dropped pronouns, and multilingual transfer learning. *Computer Speech & Language*, 89:101681.

Massimo Poesio, Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, Amir Zeldes, Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Universal anaphora: The first three years. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17087–17100, Torino, Italia. ELRA and ICCL.

Martin Popel, Zdeněk Žabokrtský, Anna Nedoluzhko, Michal Novák, and Daniel Zeman. 2021. Do UD Trees Match Mention Spans in Coreference Annotations? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3570–3576, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ondřej Pražák and Miloslav Konopík. 2022. End-to-end multilingual coreference resolution with mention head prediction. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 23–27, Gyeongju, Republic of Korea. 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*.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K.

Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye , Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, 9.

Stavros Skopeteas and Gisbert Fanselow. 2010. Focus types and argument asymmetries: A cross-linguistic study in language production. *Contrastive information structure*, pages 169–197.

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.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Annemarie Verkerk and Luigi Talamo. 2024. mini-CIEP+ : A shareable parallel corpus of prose. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 135–143, Torino, Italia. ELRA and ICCL.

Reiko Vermeulen. 2012. *The information structure of Japanese*, pages 187–216. De Gruyter Mouton, Berlin, Boston.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. SzegedKoref: A Hungarian coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lin Wang, Marcel Bastiaansen, Yufang Yang, and Peter Hagoort. 2012. Information structure influences depth of syntactic processing: Event-related potential evidence for the chomsky illusion. *PLoS One*, 7(10):e47917.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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