AURIS

Augsburg Manual for Reference and Information Structure

Compiled by Christian Chiarcos

christian.chiarcos@uni-a.de

(May 2023, version 0.1, draft, do not disseminate)

Table of Contents

1. Background and Terminology

A coherent, meaningful text can be characterized by three conditions, (semantic) consistency, (pragmatic) relevance, and cohesion (or, "connectedness"). In this manual, we focus on the annotation and subsequent analysis of the latter condition, i.e., we aim to elucidate **cohesion**, or connectedness, i.e., how each sentence is linked to an adjacent sentence in the text by means of

1. anaphoric (referring) expressions,
2. a linguistic marker for the introduction of a new topic, or
3. a semantic sentence connector ("cues").

This definition (loosely following Reinhart 1980, p.168) involves three types of analysis, i.e.,

1. the annotation of (co-)reference (what are referring expressions in the text, which entities do they refer to),
2. the annotation of topichood (what is the entity the current sentence is about), and
3. shallow discourse annotation (what are the discourse markers used, which relations do they indicate, and which utterances do they refer to)

We further limit ourselves on the first and second aspect, i.e., referring expressions and the annotation of topic continuity. Shallow discourse annotation is to be done independently, e.g., according to the schema of the Penn Discourse Treebank (Webber et al. 2019).

Coherent texts thus involve repeated mentions of the same entity as well as references to objects related in various ways to what has already been discussed, and moreover, the utterances within a coherent text are construed *about* these referents. Annotating corpora with information about such relations between elements of a text is useful both from a linguistic point of view and for applications such as information extraction.

Subsequent mentions of an entity can have the same surface form - as when the expression *the Lord Provost* is encountered twice in a text - or different ones. Anaphoric expressions are used to indicate that elements of a text are correlated. The simplest forms of anaphoric expression are used to indicate a subsequent mention of an object already introduced: typical examples of this type of anaphoric expression are pronouns such as *he* in the text *John arrived. [He] looked tired.* In the preferred reading of this text, the pronoun *he* is interpreted as an 'abbreviated reference,' to the individual John which is denoted by the expression *John*.

Besides coreference annotation itself, we include a set of linguistic features, in particular, those pertaining to information status ("givenness"), information status (here: backward-looking centers, "sentence topics") and auxiliary linguistic features (e.g., grammatical role and type of expression).

1.1 Terms

- **Coreference** is a relation between two or more textual elements, **referring expressions**, which denote the same entity. Semantically, these entities are prototypical objects or (discourse) referents.
- **Discourse referent**: an entity that is being referred to in the discourse. Note that this does not have to be a physical entity, but it can also be an imagined entity ("the unicorn … it …")
- **Anaphor**: an anaphor is a referring expression that can only be interpreted by resorting to a previously mentioned co-referential expression. The preceding co-referential expression is then referred to as **antecedent**.
- **Information Status**: The degree of prominence or familiary that a referent entertains at a certain point in discourse in the common ground (or, in the discourse model).
- **Topic**: The referent that a particular is construed about. In many cases, this is a referent that entertains a high information status, and that is repeatedly referred to, and we focus on the annotation of these "familiary topics".
- **Referential chain**: We call the series of mentions of the same referent one referential chain.
- **Information Structure**: Pragmatic structure of utterances according to the distribution of information, involving, among other aspects, information status and topichood.

- **Markable**: A (potential) referential expression that is to be annotated. Syntactically, most referring expressions are noun phrases or adpositional phrases. In the current schema, we annotate the syntactic head of the markable, only, as defined by the Universal Dependencies guidelines.

This annotation scheme is focusing on the annotation of referring expressions, i.e., nominal and pronominal anaphors and their information-structural features (information status, topichood). In addition to referring expressions, verbs may be annotated as *antecedents* of pronouns if these refer to the corresponding clause. We refer to these cases as **event anaphor**.

1.2 Referring Expressions

A *referring expression* is any linguistic form that can be used to refer to an object, person, or state of affairs (or several respectively) of the "real world" or a "conceptualized world" (as it only exists in our imagination) in a broad sense. We also include non-referring expressions, if they meet the syntactic criteria of referring expressions, e.g., "generic terms" such as in

(1) *[The whale] is known to be a mammal*

Referring expressions designate (refer to) a particular *discourse referent*, i.e., a conceptual object that representss an entity, person, or fact in the discourse model, resp., the common ground established between speaker and hearer during the discourse. A discourse referent is an abstract, conceptual object that exists regardless of whether it corresponds to an object of the world (or just of imagination).

Whether two markables are co-referent, i.e. referring to the same discourse referent, can be determined by a *substitution test*. If the substitution of anaphor and antecedent yield the same interpretation of the text, these are deemed coreferential.

1.2 Markables

Markables[1] represent spans in a text that carry one or more possible annotations, e.g., various attributes that characterize the type of the markable. We use the term *markable* for any element of the source text that is subject to annotation. Markables represent the basis for the subsequent annotation of coreference, information status, etc. This annotation scheme is limited to referring expressions, i.e. on (in the broadest sense) noun phrases, and their antecedents.

If markables they are in a coreference relation, they are given an index that indicates the referent they refer to. Coreference annotation thus consists of assignment of discourse referents to markables, represented by identifiers (mnemnonics, indixes, tags) in the `COREF` column. All corerent markables should carry the same `COREF` index.

> Note: In practical annotation, annotators should not use numbers, but a meaningful, short and unambiguous abbreviation of their own choice.

Note: As annotation is conducted here with spreadsheet software, annotators are encouraged to use the auto-complete function that such software provides. This is most effective if indexes start with different letters.

We call the series of mentions of the same referent one *referential chain*. As result of the annotation, all elements of a referential chain must carry the same index.

(2) *Susanne doesn't like [gymnastics]1, because [it]1 is very hard.*
(3) *At noon, [the Federal President]1 opened [the session]2, and in the evening, [Joachim Gauck]1 closed [it]2> again.*

The annotation task for is to process each text in reading order and identify all markables. As described below, this process is partially automated. After marking a markable, it can also be assigned various attributes that characterize the type of the markable. Here, this comprises annotations for referentiality (`REF`), coreference (`COREF`), information status (`IS`, "givenness") and backward-looking center (`CB`, "sentence topic").

These guidelines use a notation as it might be used "on paper" or in a text editor. For the practical procedure, see Sect. 2. In the examples given for illustration in this document, markables are marked by underscores (for the syntactic head), or, optionally, with square brackets [...] to clarify the boundaries of phrasal markables. Sometimes, for the sake of clarity, not all markables are marked in an example, but only those whose status is currently being discussed.

1.3 Automated Pre-Annotation

In the current workflow, automated pre-annotation will create annotations for markables, for the type of referring expressions (`NP_FORM`), their grammatical roles (`GR`), and their *possible* referentiality (`REF_AUTO`, with `?OLD` as only value so far). These annotations can be corrected by the annotator, if needed.

During annotation, dynamic pre-annotation will predict possible values for `IS` and `CB`. Again, this involves auxiliary annotations used for the automated pre-annotation of `IS` and `CB` (`GR_ANTE`: grammatical role of the antecedent, `REF_DIST`: number of sentence boundaries since last mention, `REF_DIST_ANTE`: `REF_DIST` of antecedent to *its* antecedent). These auxiliary annotations should **not** be corrected by the annotator.

1.4 Head-based Annotation

Although this manual sometimes gives phrasal markables for illustration, we only annotate their syntactic head, as defined by the Universal Dependencies (De Marneffe et al. 2021).2 As a result, markables must never overlap.

(4.a) English: *[Hans – who always had [a soft spot] [for Susanne] –] was also there.*
(4.b) German: *[Hans – der immer schon [eine Schwäche] [für Susanne] hatte –] war auch da.*

Note: Annotators should normally not need to decide which expression consistutes the head of a referring expression, as these are subject to automated pre-annotation.

1.5 About this Document

Future revisions are expected, these may include making the criteria more precise, as well as adding or amending criteria, where appropriate, or adding more examples. **However**, during an annotation campaign, these guidelines must never be changed. If an annotator feels the need for clarification or to document problematic cases, please create and provide an accompanying protocol describing the example, the problem, the decision taken for resolving or marking it in the annotation and a pointer to the data where this problem occurred. These protocols will guide subsequent revisions.

2. File Format and Editing

For the annotation of coreference and informaton structure, we use a tabular format and off-the-shelf spreadsheet software for annotation.

2.1 File Format

Following conventions for a long-standing series of shared tasks organized in conjunction with the Conference on Natural Language Learning (CoNLL) since the late 1990s, we adopt the following conventions:

- **one word per line**: One line describes one word and its annotations
- **tab-separated values**: A line consists of a fixed number of columns, separated by <TAB> (tabulator key)
- **empty line between sentences**: Two sentences are separated by an empty line.
- **use # for comments**: line starting with # are ignored when processing the file. Use this to add whatever additional information you want to express that doesn't fit the format otherwise.
- **put the text before every sentence**: To facilitate reading, the comment line before the sentence should contain the full text of the sentence.

This is the format for "raw" files, as produced by automated preprocessing. It can be opened in any text editor. For editing, you need to load these files in the spreadsheet software of your choice (see instructions below). The final deliverable should be one Excel (`*.xlsx`) file for every raw file, and it should have the same name (except for file extension) as the original file.

2.2 Raw files

The **raw files** are produced by automated pre-annotation. As part of pre-annotation, we perform tokenization (splitting words and punctuation), the detection of referring expressions, and the prediction of `?OLD` (for candidate anaphors, "primary markables") and `?NEW` (for other candidate referring expressions, "secondary markables").

Note that as part of text extracting and automated pre-annotation, some errors can occur, e.g., incorrectly split words, or incorrect type of referring expressions. **Please, do NOT fix these errors.** Instead, add a comment starting with # before the sentence. In your spreadsheet software, you might need to insert a row first.

The raw files currently contain three columns:

- `WORD`: words and punctuation characters as they occur in the text.
- `GR`: grammatical role
- `NP_FORM`: type of referring expression (noun phrase)
- `REF_AUTO`: predicted referentiality, i.e., `?OLD` or empty

2.3 Import into Spreadsheet Software: Target Files

We provide a **template file** in `*.xlsx` format that contains a number of formulas to automatize parts of the annotation. When starting with a new raw file, say, `xyz.conll` or `xyz.tsv`, make a copy of the template file and rename it such that it matches the name of the raw file, e.g., `xyz.xlsx`. We further refer to this file as your **target file**.

The template file and the target file contain the following columns:

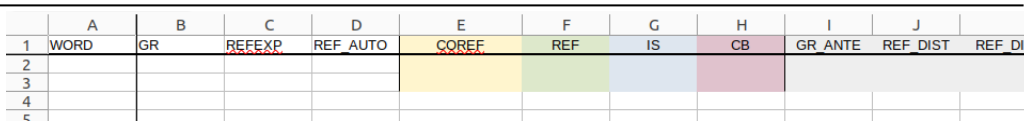
- `WORD`: words and punctuation characters as they occur in the text.
- `GR`: grammatical role
- `NP_FORM`: type of referring expression (noun phrase)
- `REF_AUTO`: predicted referentiality, i.e., `?OLD` or empty
- `COREF`: manual coreference annotation or `!!!` for an annotation to be done.
- `REF`: manual annotation for referentiality, automatically pre-annotated after `COREF` annotation.
- `IS`: manual annotation for information status ("givenness"), automatically pre-annotated after `COREF` annotation.
- `CB`: manual annotation for backward-looking center ("topic"), automatically pre-annotated after `COREF` annotation.
- the following columns (colored gray in template file) contain auxiliary annotations, these are not to be annotated, but part of the automated pre-annotation process
  - `GR_ANTE`: grammatical role of the antecedent (factor in `IS` and `CB` annotation)
  - `REF_DIST`: referential distance of the antecedent (factor in `IS` and `CB` annotation)

- REF_DIST_ANTE: referential distance annotation of the antecedent (factor in IS annotation)

Note: In the current template file, these columns are *hidden*. They will be faithfully copied if all columns (E to L) are selected as a single block before being applied to (copied and pasted into) the following annotations (see Sect. 2.4).

- COMMENT: this is a free-text column for annotators to provide information about the annotation (e.g., ambiguity), free-text comments, or pointers to more lengthy descriptions. Lengthy comments increase row height, so annotators may want to adjust column width.

Fig. 1. Template file



Open your new file xyz.xlsx in your preferred spreadsheet software. You can use any tool you like, but it **must** support reading and writing MS Excel 365 files (*.xlsx) and they **should** support Excel formulas. Possible tools include MS Office tools, LibreOffice/OpenOffice, Google Spreadsheet (in Google Docs), etc. If you have difficulties using or getting these tools, please get in touch with your instructor.

For illustration, we use OpenOffice below. Other spreadsheet software should be similar.

Now, open the "raw" file (here, xyz.tsv) in your spreadsheet software. Normally, you should be able to open it by double-clicking on it. It should open as a new table. Select the entire table and copy and paste it into your target file. Make sure not to overwrite *the first three rows* of your target file (i.e., those that contain colored columns).

Note: To select the entire table under Windows or Linux, press <CTRL>+<END> to get to the lower right corner of your data. Then, press <CTRL>+<SHIFT>+<POS1> (<CTRL>+<SHIFT>+<HOME>) to select everything until the upper left corner. Then, press <CTRL>+C to copy the entire table and <CTRL>+V to insert it at its new place. Google Docs (tested under Windows/Linux) uses Windows/Linux-style keys.

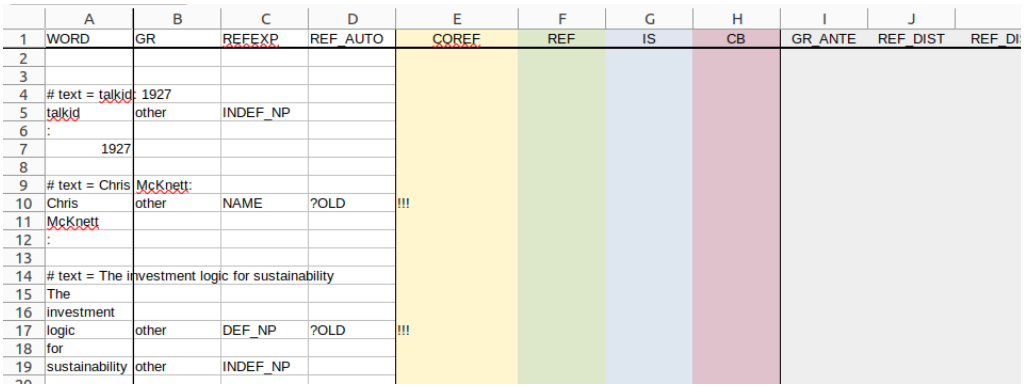
Note on MacOS: Mac keys are different. Normally, the <MAC> key should be used in place of <CTRL>.

After copying the pre-annotated data into the target file, you need to copy the *pre-annotation formulas*, too:

- Go to cell E3 (third row, column COREF). The formulas are contained in the colored and the gray columns in that row.
- Select all formulas using <CTRL>+<SHIFt>+<LEFT>, copy them with <CTRL>+C.
- Go to cell E4. Press <CTRL>+<SHIFT>+<END> to select the table from cell E4 to the end. Then, paste the formulas using <CTRL>+V. You should see colored columns for the entire text and some automated pre-annotations, now. These will update

automatically during the annotation and have to be manually corrected when needed.

Fig. 2. Target file



	A	B	C	D	E	F	G	H	I	J	
1	WORD	GR	REFEXP	REF_AUTO	COREF	REF	IS	CB	GR_ANTE	REF_DIST	REF_DI
2											
3											
4	# text = talkid: 1927										
5	talkid	other	INDEF_NP								
6	:										
7	1927										
8											
9	# text = Chris McKnett:										
10	Chris	other	NAME	?OLD	!!!						
11	McKnett										
12	:										
13											
14	# text = The investment logic for sustainability										
15	The										
16	investment										
17	logic	other	DEF_NP	?OLD	!!!						
18	for										
19	sustainability	other	INDEF_NP								

2.4 Annotation Procedure

- Annotation with spreadsheet software has a different feeling to it than just reading a text. It is highly recommended that you also look at the original plain text file, at least for a first read, before you start with with the spreadsheet annotation.

- When doing annotation, ignore headlines. For doing so, just delete the content of the `NP_FORM` and `REF_AUTO` columns for lines you identified as headlines or other pieces of metadata ("boilerplate"). For `ted-mdb.1927`, for example, this includes the following "sentences":

  - "talkid: 1927"
  - "Chris McKnett"
  - "The investment logic for sustainability"

  Note that this applies only to content you identify clearly as headline or boilerplate. If you are uncertain as to if a line is a headline or not, treat it as part of the text.

- Annotate from top to bottom, just as you read. You can use the `REF_AUTO` column for quickly jumping to the next primary markable with `<CTRL>+<DOWN>`. You can go back to the last with `<CTRL>+<UP>`.

- Alternatively, you can also go to the next referring expressing with the `NP_FORM` column.

2.4.1 COREF: Coreference

- The first requirement of the task is to assign every primary markable (`?OLD`) an ID in the `COREF` column. Every discourse referent should correspond to exactly one ID, and all co-referring expressions receive the same ID.

- If a secondary markable (annotated for `NP_FORM`, but not for `REF_AUTO`) serves as antecedent for an anaphor with `COREF` ID *x*, give it the same `COREF` ID.

    - You might want to try out for yourself if it is more convenient to either annotate all referring expressions with COREF or to only annotate ?OLD expressions and then extend this to their antecedents when needed. Please drop a note on your experiences in the annotation protocol.
- For event anaphors (e.g., if this or it refers back to a preceding clause), candidate antecedents have not been marked in the NP_FORM column. For annotating them as antecedents, select the *main verb* of the highest (in case of conjunction, first) clause you consider as antecedent. Annotate it with the same COREF ID as used for the anaphor.
    - Normally, the main verb expresses the semantic predicate of a clause or sentence, e.g., "The world is [changing] …".
    - In copula clauses, annotate the copula as antecedent, e.g., "These [are] environmental and social issues".
    - Do not annotate NP_FORM for the antecedent of an event anaphor.
    - If you have difficulties to decide which antecedent to annotate for an event anaphor, select the closest and smallest candidate, i.e., an embedded clause in favor of a main clause, the directly preceding sentence in favor of the one before, etc.
- If you encounter a non-referring ?OLD expression, delete its REF_AUTO annotation (i.e., ?OLD), but tell us which kind of non-referring expression it is using REF, etc.
- Use the COMMENT column to keep track of ambiguities or free-text comments.

2.4.2 REF: Referentiality

After annotating COREF for a referring expression, the REF column should contain the pre-annotation OLD or NEW. See the section on coreference for the meaning of these terms and other possible values.

- Please verify or revise the pre-annotation. If it is not altered, we consider it to be approved.
- If you had to delete an ?OLD pre-annotation for the current line, please annotate REF manually.
- Use the COMMENT column for comments on your annotation, e.g., to document problems.

2.4.3 IS: Information Status

After COREF and REF annotation, you will see pre-annotations for the IS column. These implement a *simplified and* ***incomplete*** *subset* of the constraints in the corresponding section that is to be manually confirmed or revised.

- Please verify or revise the pre-annotation. If it is not altered, we consider it to be approved.
- Note that the manual requires to check the applicability of annotations in a particular order. Please follow that approach here. Do **not** start with confirming the

automatically pre-annotated information status, but follow the order of statuses in the manual.

2.4.4 CB: Backward-Looking Center

After COREF annotation, you will see pre-annotations for the IS column. These implement a *simplified and* ***incomplete*** *subset* of the constraints in the corresponding section that is to be manually confirmed or revised.

- Please verify or revise the pre-annotation. If it is not altered, we consider it to be approved.
- Make sure that there is at most one CB per sentence and that all automated annotations with question marks (indicating possible CB candidates) are removed.
- By automated annotation, all referring expressions with antecedents in the last sentence are marked as CB candidates (with question marks). Make sure to remove incorrect candidates as part of your annotation.

2.4.5 COMMENT and Annotation Protocol

The COMMENT column can contain free text comments or specialized tags (e.g., for ambiguity). If you want to add more than one comment, separate them by a pipe (|).

In addition, please create an annotation protocol as an independent document to be shared along with your file. For the target file "xyz.xslx", that should be named "xyz.log" or "xyz.log.txt". Open and edit with a text editor.

Note that this view is not suited for longer text, so, longer comments should be put into the annotation protocol, but *linked* with the annotation. For doing that linking, create the comment NOTE(*abbreviation*) in the COMMENT column, using an *abbreviation* of your choice (must be unique, though, you could just use numbers). In the annotation protocol, you can then create a separate paragraph starting with "NOTE(abbreviation):" and put detailed comments there.

In addition to that, you can (and should) use the annotation protocol to keep track of any observations you made during the annotation process, e.g., difficulties in interpreting or applying the annotation manual. This will guide future revision efforts.

The annotation protocol should be saved in the same folder as the target file, and (except for the file extension), it should carry the same name.

2.5 On Evaluation

As we rely to some extent on automated pre-annotation, we need to quantify the number of average revisions of pre-annotated values per file and annotator.

3. Automated Pre-Annotation of Markables

Texts for annotation should be automatically pre-annotated for referring expressions. This document contains the guidelines for the algorithm. Normally, this is irrelevant for manual annotation and can be skipped by annotators.

We provide a pre-annotation routine that identifies referring expressions along with

- their morphosyntactic type (NP_FORM)
- for potential anaphors, their expected referentiality (REF_AUTO, only value is currently ?OLD), and
- their grammatical role (GR).

We describe NP_FORM and REF_AUTO as part of markable identification. GR annotation is described separately.

3.1 Types of Markables

The annotation task is to process each text in reading order and annotate/verify all markables (automatically pre-annotated) and their antecedents (including cases in which these are not pre-annotated).

The scheme distinguishes between primary and secondary markables. Primary markables are *always* subject to annotation. Secondary markables are only annotated if they happen to serve as antecedents for primary markables. In earlier versions of this schema, explicit annotations for primary and secondary markables were included. This is, however, not necessary, as they are merely a technical device to guide the annotation process.

More specifically,

- **primary markables** (PM, pre-annotated for REF_AUTO as ?OLD, and for NP_FORM) are candidate anaphors, i.e., noun phrases whose grammatical features suggest that their discourse referent is or could be identifiable by the hearer.[1] For German and English, these are definite NPs and pronouns. For languages without grammatical marking of definiteness, these are all nominals and pronouns.

  Primary markables are automatically extracted. The task of annotation is to assign every primary markable either an antecedent or a flag that marks them as new or non-referential.

- **secondary markables** (SM, pre-annotated for NP_FORM, but not REF_AUTO) are antecedents for anaphoric expressions which have not been detected as primary markable.

  Typical secondary markables are indefinite expressions (NP with indefinite article, *a dog*) or without article (*good weather*).

In automated pre-annotation, all primary markables are assigned the referentiality (REF_AUTO) value ?OLD. Secondary markables are annmotated for their NP_FORM, but not

for REF_AUTO. The task of manual annotation is defined as annotating all primary markables and their antecedent from beginning to end, so that the presence of ?OLD indicates that a text has not been fully annotated.

Word forms that are confirmed to be syntactically bound are not to be annotated. Forms that are ambiguous between bound and anaphoric pronouns are annotated as primary markables and to be disambiguated manually.

For every text,

- primary markables represent the set of all candidate anaphors,
- primary and secondary markables represent the set of all candidate referring expressions

Candidate antecedents may also be outside this set, and will not be automatically annotated, in particular, event anaphor.

3.2 Identifying the Syntactic Head

We only annotate the syntactic heads of markables according to the Universal Depenency syntax. Thus, markables must never overlap:

> (1.a) English: *[Hans – who always had [a soft spot] [for Susanne] –] was also there.*
>
> (1.b) German: *[Hans – der immer schon [eine Schwäche] [für Susanne] hatte –] war auch da.*

This also entails that referring expressions can *only* be annotated if they are identified as independent words by the word segmentation procedure adopted for that particular language. In (1.c), *Denver* and *bancruptcy* can only be identified as markables if they are (automatically annotated as) independent tokens.

> (1.c) *The [Denver]?-based concern, which emerged from bancruptcy ... its new, post-[bancruptcy]? law structure ..."* (WSJ, 1328)

When converting head-based annotation to span-based annotation in downstream tasks, we assume that all dependents of a syntactic head are to be included in the markable:

> (2.a) *[This right]right may not be invoked [in the case of prosecutions arising from acts contrary [to the purposes [of the United Nations]UN]purp]prosec.* (www.unhchr.ch/udhr, shortened)
>
> (2.b) *[Dieses Recht]right kann nicht in Anspruch genommen werden [im Falle einer Strafverfolgung auf Grund von Handlungen, die [gegen die Ziele [der Vereinten Nationen]UN]purp verstoßen]prosec.* (German, www.unhchr.ch/udhr, shortened)
>
> (2.c) *[Это право]right не может быть использовано [в случае преследования, основанного на совершении деяния, противоречащего [целям [Организации Объединенных Наций]UN]purp]prosec.* (Russian, www.unhchr.ch/udhr, shortened)

3.3 Primary Markables (REF_AUTO=?OLD)

Primary markables are automatically extracted from a syntactic analysis.[2] The following criteria define the algorithm. Normally, annotators do not have to annotate PMs and they can skip this section. However, if you feel there may be an anaphoric expression that was missed in automated extraction, please resort to these definitions.

> Note: Incorrect PM prediction can result from parser errors. Annotators should mark manually introduced PMs with the comment `manual PM`.
>
> Note: Annotators must never delete an incorrectly extracted PM annotation, but you can mark it as non-referential and add the comment `parser error` to the annotation.

3.3.1 Pronouns (NP_TYPE=pron)

Pronouns include personal pronouns, demonstrative pronouns, pronominal adverbs, and possessived pronouns and *both* in nominal use (i.e. not as a determiner),[3] e.g.,

(3) *[I] saw [her] yesterday.*

If automated pre-annotation operates on a language/annotation schema that doesn't distinguish these types of pronouns from other (non-referring) types of pronouns, *every* pronoun should be annotated as primary markable.

> Note: Interrogative pronouns are not primary markables, but can serve as secondary markables.
>
> Note: Relative and reflexive pronouns are not primary markables, if annotated by pre-annotation, they should be annotated as REF=BOUND. Their automated NP_TYPE annotation can be deleted.

3.3.1.1 Personal Pronouns (NP_TYPE=pron.pper)

Personal pronouns include (the language-specific counterparts of) English *I, me, you, he, him, she, her, it, we, us, they, them*.

Note that so-called "generic pronouns" (*we, you, they*, in German *wir, du, sie* (without specific reference), *man, einer*) are considered as indefinite, but that they cannot be automatically identified. Thus, they are annotated as primary markables.

> Note: Reflexive pronouns (English *herself*, etc.) are not PM. Pronouns that are formally ambiguous as to whether they are reflexive or personal pronouns (e.g., German *mich* "me; myself"), are PM, and should be manually marked as REF=BOUND in the annotation.
>
> Note: Other non-referring pronouns (e.g., expletive *it* or generic *you* in the sense of "anyone") are likewise not to be deleted but to be annotated manually.

3.3.1.2 Possessive Pronouns (NP_TYPE=pron.ppos)

Possessive pronouns include (the language-specific counterparts of) English *my, mine, your, yours, ….*

3.3.1.3 Demonstrative Pronouns (NP_TYPE=pron.pds)

Demonstrative pronouns occur with two optional sub-classes:

- NP_TYPE=pron.pds-prox: proximal *this, these*, *this one*, German *der, die, das, …*
- NP_TYPE=pron.pds-dist: distal *that, those*, *that one*, German *dieser, diese, dies(es), jener, jene, jenes, derjenige,* and the like.

Note that demonstrative pronouns *such*, in German *solch*, are considered indefinite.

> Note: Relative pronouns (English *which*, etc.) are not PM. Pronouns that are formally ambiguous as to whether they are relative or demonstrative pronouns (e.g., English *that* in relative clauses), are PM, and should be manually marked as REF=BOUND in the annotation.

3.3.1.4 Pronominal Adverbs (NP_TYPE=pron.padv)

Pronominal adverbs are derived from pronouns but grammaticalized as adverbs. If pronominal adverbs can still be interpreted as / replaced by a referring expression in a particular language, they should be included as primary markables. However, we exclude references to time and place of the speaker (*here*, *hence*) if these are unambiguous in their deictic function, as well as interrogative adverbs (*where*, etc.).

Examples: Pronominal adverbs in German include *da* “there, then”, *dort* “there”, *daneben* “next to it”, *dahin* “(towards) there”), *davor* “in front of that; before that”, or *deswegen* “because of that”.

> Note on English: Normally, pronominal adverbs are not recognized as referring expressions in English, but they can indeed be substituted with prepositional phrases. For English, we annotate adverbs (starting with) *there* (unless expletive) and *thence*, e.g., *there*, *thereafter*, *therefore*, *thence*, *thenceforth*. We exclude the analoguous *here* and *hence* because they are exclusively deictic, not anaphoric, whereas there and thence could also have an anaphoric function (*therefore ~ for this reason*, *thence ~ from there*).

3.3.2 Definite Descriptions (NP_TYPE=def-np)

A description (NP or PP) is definite if it contains the determiner *both*, a demonstrative or possessive pronoun or a genitive attribution. Optionally, this can be made explicit with sub-types.

3.3.2.1 With Demonstrative Determiner (NP_TYPE=def-np.dem)

(4) *[that pizza]*, *[this pizza]*

Demonstrative NPs involve optional differences with respect to their relative proximity, with optional subtypes

- `NP_TYPE=def-np.dem-prox`: proximal *this man*, ...
- `NP_TYPE=def-np.dem-dist`: distal *that man*, ...

3.3.2.2 With Possessive Modifier (`NP_TYPE=def-np.poss`)

Constructions with possessive pronouns.

(5) *[his pizza]*

Also includes potentially genitive or possessive modifier, if these are (potentially) anaphoric

(6.a) *[John's pizza]*

(6.b) *[the pizza of John]*

(6.c) *[the other man's pizza]*

(6.d) *[this man's pizza]*

but not: *[a man's pizza]*

3.3.2.3 Quantified Definite NP (`NP_TYPE=def-np.quant`)

At the moment, this includes cases where a quantifier is combined with a definite article (`the two men`) or with determiner 'both'

(7.a) *[the two pizzas]* (7.b) *[both pizzas]*

But not: *two pizzas*. As for constructions like *two of these pizzas*, this is formally a possessive construction.

Note: Stede et al. (2016) include *all*+NP here. needs to be double-checked.

3.3.2.4 With Definite Article (`NP_TYPE=def-np.the`)

Any NP with a definite article not covered by any aforementioned def-np category

(8) *[the pizza]*

3.3.2.5 NP with "Other" (`NP_TYPE=def-np.other`)

Definite NPs containing adjectives like *other*

(9) *the other man*

Note that the `other` flag can be attached after *any* def-np subtype, so, the following tags are valid:

- `def-np.other` (for otherwise unclassified definite NPs)
- `def-np.dem.other` (for otherwise unclassified demonstrative NPs)

- `def-np.dem.prox.other`: *this other man*
- `def-np.dem.dist.other`: *that other man*
- `def-np.poss.other`: *his other goal*
- `def-np.quant.other`: *the two other guys*
- `def-np.the.other`: *the other man*

3.3.3 Proper Names and Titles (NP_TYPE=ne)

Typical instances of proper names are geographic places (*Philadelphia*), persons (*Judge Jenkins*), companies (*Morgan Stanley & Co.*), newspaper titles (*The New York Times*), political, social or financial institution names (*Congress, European Investment Bank*). Proper names can include noun modifiers or be heads of a definite or indefinite description. In this case, the whole description has to be marked up, not just the head.

(10.a) *[Bertolt Brecht]* (full name)

(10.b) *[Bert Brecht]* (reduced full name)

(10.c) *Brecht* (surname)

(10.d) *Bertolt* (first name)

(10.e) *Bert* (nickname)

(10.f) *BB* (abbreviation)

(10.g) *the well-known Brecht* (name, modified by a definite description)

(10.h) *Brecht, who is author of the "Dreigroschenoper"* (proper name + clause)

(10.i) *Brecht, author of the "Dreigroschenoper"* (proper name + apposition)

Complex proper names are only treated as a single markable and are not further divided. If the internal dependency structure is transparent, annotate the syntactic head. For names composed of given and family names, we consider the name of the individual to be head, and the name of the family as modifier. If the structure of a name is not transparent to a common speaker of the language, annotate the first word that is not clearly recognizable as a modifier.

(10.j) [Dr. Mueller]

(10.k) [Dr. Martin Luther King, Jr.]

(10.l) [Prince Dipangkorn Rasmijoti Sirivibulyarajakumar of Thailand]

(10.m) [Heidelberger Druckmaschinen Vertrieb Deutschland GmbH]

Standalone titles that can stand in for an individual (*Mr./Ms./Dr./President/Chairman*) are treated like proper names, e.g.,

(11.a) *Schröder1...Fischer2 ... Die anfängliche Überreaktion von Kanzler1 und Außenminister2...*

In (11.a), *Kanzler* and *Außenminister* have to be annotated as primary markables, because proper names are inherently definite

Parts of complex proper names cannot be analyzed separately. So, in the following example, *Petrie* in *[of Petrie Stores Corp.]* should not be annotated!

(11.b) *[Milton Petrie, chairman [of Petrie Stores Corp.] said...*

3.4 Secondary Markables (no `REF_AUTO` annotation)

Every nominal phrase or pronoun which is neither primary markable nor (confirmed to be) syntactically bound, is subject to automated pre-annotation. Secondary markables are referring expressions that are unlikely/impossible anaphors, but that could *introduce* new discourse referents.

> Note: At the moment, these are automatically annotated for `NP_FORM`, but not for referentiality (`REF_AUTO`).

Common types of secondary marakbles include: indefinite NPs and indefinite or non-referring pronouns.

Annotate the secondary markable only if you are certain about the reference. If another reading is equally possible or feels more likely, do not annotate the secondary markable. (Add a comment about your uncertainty.)

(12) *I saw [a cat] tonight in the street. It(= the cat) was gray.*

but not: *I saw a cat tonight in the street. It(= the night/expletive?) was pitch black.*

3.4.1 Indefinite NPs (`NP_TYPE=indef-np`)

With optional sub-types:

- `NP_TYPE=indef-np.a`: NP with an indefinite article, e.g., *a fox*:

  (12.a) *There is a [a fox] running across the street. It's fast!.*

  (12.b) *I last saw [a fox] about three years ago! It came from the forest.*

  Also includes indefinites with other, e.g. *another man*

- `NP_TYPE=indef-np.quant`: NPs with indefinite quantifier, also including quantified expressions not otherwise annotated as primary markables

  (13.a) *[some people]*

  (13.b) *[some plants]*

  (14.c) *[thirty grams]*, *[two companies]* (quantified indefinite NP)

  Borderline case: indefinite NPs with an article that is identical (or at least derived from) the cardinal number *one* should be considered as quantified iff. a corresponding set of individuals has been previously evoked and the membership relation marked as being relevant.

For English, the latter condition should hold for *one*, but not for *an, a*, for German, the membership relation should be regarded as being prominent if a substitution of the indefinite article *ein, eine* by colloquial *'n, 'ne* appears to be unlikely.

- `NP_TYPE=indef-np.bare`: articleless NP, especially "bare plurals", but also singular expressions.

  (14.a) *I have eaten [cookies]SM* (bare plural)

  (14.b) *Today will be [good weather]SM* (bare singular)

3.4.2 Non-anaphoric pronouns (`NP_TYPE=pron`)

With optional sub-types:

- `NP_TYPE=pron.pint`: interrogative pronouns: *who, where, when, ...*
- `NP_TYPE=pron.pds`: indefinite demonstrative pronouns, e.g., *such*, in German *solch*
- `NP_TYPE=pron.pind`: indefinite pronouns, e.g., *somebody*, or German *man*. Also includes pronominal indefinite quantifiers, e.g., *some* in *some of that*.

3.4.3 Other expressions (`NP_TYPE=other`)

We consider every syntactic argument of a verb to be a potentially referring expression. If not matched by any of the aforementioned conditions, we treat verbal arguments as secondary markables. This can happen if an argument is a foreign language expression that is not assigned proper POS tags, but instead just marked as foreign (e.g., `X` in Universal Dependencies). Note that the annotation of referentiality for `other` nominals is tentative, only.

3.5 Automated Pre-Annotation

- Check every nominal, pronoun, prepositional phrase, or proper name
- If it is a primary markable, pre-annotate it with referentiality `?OLD`
- If it is a secondary markable, do not annotate it for referentiality.

Note: Alternatively, annotate secondary markables with referentiality `?NEW`.

For every markable, annotate the syntactic head (as defined by the Universal Dependencies, exceptions as mentioned above) for type of referring expression.

3.6 Trouble-Shooting

3.6.1 Demonstratives

NPs with demonstrative determiner (English *this NP*, *that NP*, German *diese NP*, *jene NP*), and demonstrative pronouns (German *dieser*, English *this*, *that* [if not used as relative pronoun]) are primary markables.

The demonstrative pronoun *such* (German *solch*) is considered as indefinite. Refererring expressions with *such* should not be annotated as primary markables.

3.6.2 Bound Pronouns

Do not annotate bound pronoun, if these can be identified on grounds of their form or annotations. If a pronoun is ambiguous in its surface form and cannot be unambiguously confirmed as bound pronoun from the syntactic annotation, treat it like a primary markable and annotate with referentiality `?OLD`. Only in these cases, the annotator should then annotate referentiality `BOUND`.

3.6.3 Treatment of Quantifiers

Quantified NPs *(some of them, all the members)* are annotated as either definite or indefinite, whereas each case has to be considered individually. Substitution test: *all days* –→ *all these days* –→ definite. In automated pre-annotation, every nominal phrase whose form does not rule out a definite interpretation should be treated as primary markable.

We regard NPs with certain quantifiers in determiner position such as *both* as definite, since German *beide* as English *both* normally presupposes the existence of exactly two discourse-old entities (cf. Zifonun et al. 1997).

Both in nominal use is annotated as a personal pronoun.

3.6.4 Possessive NPs

Primary markables include

(15.a) *his house* (15.b) *the old man's house*

Note that the possessor must be a primary markable, too: *[[his] house]*, *[[the old man's] house]*.

Descriptions with a genitive attribution are regarded as possessive iff. a definite genitive attribution replaces the determiner, except for *of* -constructions (cf. ex. 16.b).

(16.a) *the old man's house* (definite possessive description, cf. *his house* and **the the old man's house*) (16.b) *the house of the old man* (definite non-possessive description: determiner in *the house* relates to the house itself and not to its possessor) (16.c) *an old's man house*(this is not a primary markable - the possessor is indefinite)

Possessive NPs with indefinite possessor (16.c) are secondary markables. However, possessives with proper names should generally be considered as primary markables.

(16.d) *in US efforts*

This is a primary markable because there is a reading, where the phrase could be replaced with *in the US efforts*.

3.6.5 Appositions

Appositions are treated like predications. That is, they serve neither as antecedents nor anaphors. So, in the following example, *chairman* in *chairman of Petrie Stores Corp.* should not be annotated!

(17) *[Milton Petrie, chairman [of Petrie Stores Corp.] said...*

3.6.6 Stranded Quantifiers

An NP can be incomplete by elision and, at first glance, not meet the criteria of a markable. For example, individual numerals are not usually PM, but if their head noun is elided, they serve as heads of NPs, they can require an antecedent.

(18) *Now only three of the 12 judges - [[Pauline Newman]n, ([Chief Judge Howard T. Markey, 68]m)two*1*, and ([Giles Rich, 85]r)two* 2 *- have patent law backgrounds]. [The latter two]two and [Judge Daniel M. Friedman, 73]f , are approaching senior status or retirement.* (WSJ corpus)

As these cases cannot be automatically identified, all pronominal numerals are to be annotated as primary markables.

(18') *Ich hatte [zwei Stunden]PM eingeplant, aber es wurden letzlich [drei]SM.* (German) (18") *I had planned for [two hours], but in the end, it was [three]SM* (English)

3.6.7 Proper Noun vs. definite NP

Note that if a proper noun is not a head of an NP, the NP is annotated as definite or indefinite respectively.

(19) *the river Yukon*

In (19), *Yukon* is the head. *Yukon* is a proper name, so the whole phrase is annotated as proper name.

(20) *the Yukon office*

In (20), *office* is the head, *office* is not a proper name, so *the Yukon office* has to be annotated as a definite NP.

3.6.8 Non-referring Primary Markables

Non-referring markables (NM) are primary markables whose function is *not* to refer to a discourse referent. Non-referring markables are to be *manually* given the appropriate referentiality value in subsequent annotation (`GEN`, `EXPL`, `PRED`. `IDIOM` or `other`, see there). For automated extraction, they are treated like primary markables.

3.6.9 Do NOT annotate

- expletive expressions

(21) *Then, when it would have been easier to resist them, nothing was done* (expletive *it*).

- *Es*-pronouns, pronominal adverbs, which are controllers of relative clauses

(22) *Dazu kommt, dass in Werder am 24. Februar ein Bu¨rgermeister gewa¨hlt wird und es bisher als sicher galt, dass CDU-Amtsinhaber Werner Gr¨oße unangefochten bleibt.*

Dazu...dass, es...dass should not be annotated as markables (*Dazu* and *es* are controllers of relative clauses).

- pronominal adverbs functioning as discourse markers

(23.a) *Ich habe dich angesprochen, damit du mir zuhörst*. "I am talking to you to let you know that you must listen to me." (23.b) *Ich habe dir das gesagt, damit du weißt, dass du mir zuh¨oren sollst*. (23.c) *Ich habe dir das gesagt, dass du weißt, dass du mir zuh¨oren sollst.*

- relative pronouns

Relative pronouns are annotated together with the whole relative clause it triggers as one single markable (cf. *[The car that went through his garden wall]...*). If a form cannot be unambiguously classified as a relative pronoun, apply the following test: it is a relative pronoun if it can be substituted by "which" respectively "welch" in German. However, relative pronouns in possessive constructions (i.e. for which the test for relative pronouns fails) are annotated as possessive pronouns (see possessive NPs, p. 10).

(24) *Und so schielen die Israelis nach Washington, an dessen /*welchem Tropf sie wirtschaftlich und militärisch hängen,...* (24') *Und so schielen die Israelis nach Washington, das/welches sie wirtschaftlich und militärisch unterstützt* (*das* is a relative pronoun).

Alternatively, the following test can be applied: substitute a pronoun in question with a possessive construction. If it works, you have a possessive pronoun, not a relative one.

(25) *die Frau und deren Kinder = die Frau und ihre Kinder*

The annotation is as follows in this case:

(26) *Und so schielen [die Israelis]i [(nach Washington)w, [an [dessen]w Tropf] [sie]i wirtschaftlich und milit¨arisch h¨angen]w' ,...*

- prepositional phrases with prepositions *as, than*, *bis, als, wie* (in German) Such prases are annotated as normal NPs, i.e. *bis* and *als* are not included. [3]

- nominal premodifiers in compound nouns

(27) *peanut butter, airline analyst, the creditors commettee, investment bank*

Peanut, airline, cretitors and *investment* are no separate markables. Note that in *the creditor's opinion*, *the creditor's* is annotated as a markable, since it is a nominal in genitive and thus not a part of a compounds.

3.6.10 Idioms and Collocations

Primary markables in idioms and collocations, if identifiable in automated pre-annotation.

(28) *It sent Kate into the pits when she learned from her "friend" Martha, who seemed to get off on laying bad trips on people, that Harvey was getting it on with Carol.* [Gib94, p.265]

According to Gibbs, we find several idiomatic phrases in this example, some of which contain pronouns or full NPs – potential primary markables.

However, they should not be annotated as such, e.g. *into the pits* meaning "to be depressed", *get it on* meaning "having sexual relations", neither *the pits* nor *it* can be referred to.

Note that we consider only *conventionalized* idiomatic expressions as idioms in our sense,

i.e. markables within productive metaphors are annotated as usual, e.g. *das schlingernde City-Schiff City-Schiff* - a metaphor that occurred and can only be understood with respect to a specific text.

3.7 Grammatical role annotation (GR)

Grammatical role annotation is extrapolated from (automated) annotation according to Universal Dependencies conventions (either UD v.1 or v.2), with the following rules:

- `SBJ`: nominal subject (UD edge: starts with `nsubj`)
- `OBJ`: grammatical object (UD edge: contains `obj`)

  Note: Chiarcos and Krasavina (2005) distinguished indirect and direct objects. However, within the UD community, the notion of indirect object has been criticized, and the usage of `iobj` seems to be inconsistent.
- `other`: every other referring expression is annotated `other` (i.e., every element that carries `NP_FORM` annotation but has not been assigned a `GR` annotation before, includes both primary and secondary markables, but not antecedents of event anaphors)
- for referring expressions in dependent clauses or adnominal constructions, we append the depth of syntactic embedding as a numerical suffix (i.e., replace existing `GR` annotation `$gr` with `$gr+_depth`, e.g., `SBJ_2` for the subject a relative clause directly depending on the main clause. Embedding depth is calculated over UD trees.)

  Note: In Chiarcos and Krasavina (2005), these were included under `other`

Note: According to these rules, nominals that are not integrated into the clausal structure are considered as `other`.

These rules are implemented in `sparql/gr.sparql`.

Note: TODO: update SPARQL script for new abbreviations

4. Nominal coreference

We annotate referential chains by co-indexing all referring expressions that refer to the same referent.[1]

4.1 Scope and Aim of Annotation

1. assign every referential expressing an index that unambiguously identifies its discourse referent
2. annotate antecedents of anaphoric expressions accordingly (regardless of whether these are referring expressions or not)
3. annotate all remaining nominal and pronominal expressions as non-referential

Noun phrases, names and pronouns are automatically pre-annotated as markables.

We annotate every markable with - `COREF`: an abbreviation ("index") for the discourse referent (or its absence), - `REF`: the type of reference (see below), and - `COMMENT`: optional structured or free-text comment indicating uncertainties or design decisions. Note that this column is used for both coreference and following annotations

We do not consider syntactically bound expressions as coreferential with their controller (e.g., predicative nominals in copula sentences, or relative pronouns), as the relationship with the hypothetical antecedent is expressed by syntactic means.

4.2 Annotation Procedure

Annotate all referents in the order that they occur in the text with

1. `COREF` (discourse referent tag): which referent a referring expression refers to, or `_` if the expression is not referential

2. `REF` (referentiality): reference type as defined below

3. `COMMENT` (optional comment): can contain free-text comments as well as ambiguity annotations (see below).

In spreadsheet-based annotations, some of the values for refentiality are automatically suggested (`REF_AUTO`). This may help your decision for the annotation of `REF`, but please make sure to verify that properly.

Follow the following steps (and see below for additional instructions on the annotation of `REF`):

- If the markable introduces a new discourse referent, annotate it with a new index and the reference type `NEW` (or `SIT` for first person, second person or dates). This the default for indefinites and bare nouns.
- If a markable refers to previously introduced referent, annotate it with the index used for the antecedent and annotate its reference type.
- If a markable is not a referring expression, annotate it with an empty index (`_`) and annotate the type of non-reference.
- If a markable refers to two (or more) distinct, previously mentioned discourse refents ("group reference"), create a new index for the group, followed by `\>` and the comma-separated indices of all discourse referents. Assign it the reference type `GROUP`.

Use the `COMMENT` column to annotate ambiguity and provide comments as needed. If you need to come back to a passage to confirm your annotation, put that into comment, as well.

For every referent, after annotating discourse referent index and referentiality, annotate information status. In the spreadsheet, this is partially automated, but need to be confirmed.

4.3 Coreference (`COREF`)

For annotating coreference, use user-defined abbreviation/mnemnonic/tag that indicates which referent a referring expression refers to, or _ if the expression is not referential.

> Note: The discourse referent tag, or "index", is an abbreviation that a user should chose at the first mention of the referent. Chains of antecedents and anaphors should all have the same index.
>
> Note: In this document, we use numerical indexes for presentational reasons. In annotation, create abbreviations/mnemnonics as you see fit. Numerical indices are discouraged in actual annotation.

4.3.1 Substitution Test

A replacement test can be used to check whether a referential expression e belongs to a chain k: If it is true for every noun s (noun, proper noun) in k, that the replacement of e by s changes the interpretation of the text is not changed, then e belongs to chain k and a coreference relation to the last element of the chain is to be annotated.

According to this scheme, we only annotate coreference relations that express a real identity between discourse objects. A "semantically loose" connection between a definite NP and another nominal is therefore not sufficient. For this purpose the following

Test: To find out if two nominal descriptions are coreferent, try to substitute them with each other. (Certain transformations may be necessary, such as removing prepositions from markables.) Note that every previous coreferent markable has to be compatible with this substitution as well.

Note that this test has some issues with metonymy, i.e., substituting a word for another word closely associated with it. Cases of metonymy in text should be annotated as coreferent if and only if the subsitution test holds for all coreferring nominals: *the State Department said... - the State Department officials claimed....*

(1) *Als 1999 die im Rahmen der Dorferneuerung neu gestaltete [Radeweger]1 Ablage inklusive Seebrücke mit viel Pomp eingeweiht wurde... Doch mit der Nachrüstung tut sich [Radewege]1 schwer ... Zu teuer, zu hässlich sei die Anlage, sagen die Meinungsführer [im Gemeinderat]?* (maz-6488)

In (1), *[Gemeinderat]* could be considered coreferent with *[Radewege]1*. Yet, although both are exchangeable by means of metonymy, the substitution test fails for *[Radewege]1* , since *neu gestaltete Gemeinderatsablage* is not appropriate *in that context*. Accordingly, *[Gemeinderat]* should receive a separate index.

4.3.2 Event Anaphor

Pronominal event anaphors are annotated along with their antecents.

The antecedent of an event anaphor is normally a sentence, a clause or verb phrase. If so, select the main (lexical) verb as antecedent. Hence, in this example, we annotate *gewonnen* as we encounter the referring expression *das* “this”.

(2) *Gestern hat Bayern München schon wieder gewonnen1. [Das]1 hat Jan ziemlich gestört. Marianne hingegen war [davon]1 begeistert.* (Non-event anaphors skipped.)

Note that antecedents of event anaphors are not automatically pre-annotated but have to be manually created.

4.4 Ambiguity

Ambiguity can be annotated on demand in the `COMMENT` column.

4.4.1 Dealing with Ambiguous Antecedents

The assignment of an antecedent will be fairly straightforward in most cases. However, it is possible that several interpretations are *equally* plausible in the eyes of the annotator. Consider ex. (3):

(3) *Je kleiner die Kicker2,OLD,AMBIG:COREF(2,1) daherkommen, desto größer wird der Gegner1,OLD,AMBIG:COREF(1,2) geredet...* (German, maz-10374) “The smaller the kickers appear, the greater [the rivals]d?/u? are rumoured to be.” (PCC, 10374)

Antecedent of die Kicker “kickers” depends on the understanding of the “size” metaphor, it can be either the Ukrainian team (presented as having short players), or the German team (which has not been favored in the first match), or a generic description (which would mean that the sentence is not directly linked with the discourse).

If different interpretations are equally possible, apply the following disambiguation preferences in the following order:

- prefer anaphoric interpretation to antecedentless one
- for antecedents, prefer a primary markable (REF_AUTO=?OLD, REF=OLD, etc.) over a secondary markable (no REF_AUTO, REF=NEW, etc.) or a group reference
- prefer a discourse referent that is more frequently mentioned in preceding discourse
- if several discourse referents are equally frequent, prefer the last mentioned discourse referent

For the example, the generic reading is excluded by the first preference. However, we still have the choice between two possible antecedents. The substitution test (see above) fails to determine a unique antecedent, as both possible substitutions are plausible, depending on whether "size" refers to physical size or anticipated defeat. The second and third criteria are designed to produce longer anaphoric chains. They result in a preference for the German team as the antecedent of die Kicker.

In annotation, then, **mark the ambiguity** (in COMMENT)

4.4.2 Types of Ambiguity

Ambiguity is to be annotated in the COMMENT field, using pre-defined tags (if applicable) or plain text descriptions (otherwise). Optionally, ambiguity tags can be followed by a more detailed description in round parentheses (...).

The following tags can be used to mark ambiguous

1. AMBIG:COREF (ambiguous antecedent):7 There is uncertainty as to which is the "right" antecedent for an anaphor (or, controller for a cataphor). See above for antecedent selection preferences, provide referent index for all equally likely antecedents in round parentheses

   (4) *In a letter, [prosecutors]p told [Mr. Antar's lawyers]l that because of the recent Supreme Court rulings, [they]p/l? could expect that any fees collected from Mr. Antar may be seized.*

2. AMBIG:REL: There is uncertainty as to whether an anaphoric relation exists or which type it is (anaphoric vs. bridging or event, i.e. contextual inference)

   This is sometimes the case with definite NPs. In the example below: If it is unclear whether the *confrontation* is identical to the *conflict*, the coreference should be annotated and the markable should be marked with this attribute. It is not necessary to provide a more detailed description.

   (5) *This conflict is ... Therefore, the confrontation ...*

3. AMBIG:IDIOM: There is uncertainty as to whether a markable could be understood as a referential expression or as part of an idiom. Annotate anaphoric reading and mark the ambiguity.

4. AMBIG:EXPL: There is uncertainty as to whether a pronoun is an expletive (and therefore non-referring) or whether it is anaphoric. Annotate the anaphoric relations and mark the ambiguity. No description necessary.

   (6) *At stake was an $80,000 settlement involving who should pay what share of cleanup costs at the site of a former gas station, where underground fuel tanks had leaked and contaminated the soil. And the lawyers were just as eager as the judge to wrap [it] up.*

   It can either be interpreted as referring to *an $80,000 settlement* or as a part of a lexicalized expression *to wrap it up* where *it* does not have any particular reference.

   This can be made clearer with a constructed example:

   (7.a) *She looks out of the window. ItEXPL is dark.* (expletive)

   (7.b) *Your cat1 has a nice color. It1 is dark, much more so than mine.* (anaphoric)

   (7.c) *The cat1 is hard to see. It1,AMBIG:EXPL is dark.* (ambiguous)

5. AMBIG:COREF_REL: There is ambiguity with respect to both antecedent and relation

   (8) "There seems to be a move around the world to deregulate the genera- tion of electricity," Mr. Richardson said, and Canadian Utilities hopes to capitalize on it.

   On it refers either to *a move around the world to deregulate the generation of electricity*, or to the whole clause beginning with *there* and ending with *electricity* (event anaphora).

6. AMBIG:other: other cases of ambiguity. Please provide a description in round parentheses.

If more than one kind of ambiguity applies, e.g., both ambiguity of antecedent and ambiguity of an anaphoric relation, then provide all of the corresponding tags (and descriptions), separated by comma.

4.5 Referentiality (REF)

Every markable that is not assigned an antecedent is to be annotated for referentiality.

In spreadsheet-based annotations, some of the values are automatically suggested in REF_AUTO. Please make sure to verify all of them. Automated pre-annotation generates the value ?OLD for all candidate anaphors ("primary markables") and, optionally, ?NEW for all other candidate referring expressions ("secondary markables").

> **To be confirmed**: These values need to be manually replaced by the annotator. An annotation project that still contains ?OLD or ?NEW annotations will be considered incomplete and must not be further processed.

1. OLD: A unit of discourse that can be interpreted based on the preceding context ("discourse-old").2

2. NEW: Discourse entity mentioned for the first time. This includes referents that can be inferred by the hearer ("discourse-new, but hearer-old"). 3

3. CAT: Discourse cataphor, i.e., reference to a new entity introduced into the discourse with an underspecified nominal expression whose exact denotation becomes clear only from subsequent descriptions. This "scene-setting" effect is a rhetorical device employed to engage readers in literary texts.4

   When reading the text of the annotation example below, it is initially unclear what *Fußball-Weltmacht* and *Winzling* refer to. This becomes clear only in the next sentence, when the German team and Ukraine are mentioned.

   Note: Syntactic cataphors are not included here. Syntactically bound pronominal cataphors are annotated as BOUND.

4. GROUP: Referring expressions that designate groups can serve as antecedents of nominal markables and can be annotated as a group. 5

   (9.a) *[Montedison]1 now owns about 72% of [Erbamont's]2 shares outstanding.*

   (9.b) *[The companies]3>1,2 said ... a sale of all of [Erbamont's]2 assets ... [to Montedison]1 ...*

   (9.c) *[The companies]3 said ...* (WSJ, 660)

   Note that the second reference to *the companies* in (9.c) is annotated as a **plain anaphoric reference** to the established group, not as a group reference to the individual companies mentioned in the meantime.

5. BOUND: Pronouns that are syntactically bound, e.g. reflexive pronouns. Also, possessive pronouns governed by nominal expressions in the same sentence are annotated as BOUND, cf. in (8.b) below *Mit seinem3,BOUND Tor*.6

   Notes: Reflexive pronouns (which are obligatorily bound) are not annotated as markables if they can be identified on grounds of their form (e.g., English *himself*, German *sich*). Only if a form is ambiguous between a reflexive and pronominal reading (e.g., German *mich*), reflexive pronouns are annotated as BOUND.

6. SIT: situationally evoked. In written texts, this applies only to first and second person, non-reflexive pronouns and to temporal expressions.5

   Note: SIT is to be annotated at the first mention, only, subsequent references to the same entity are to be annotated as OLD.

7. GEN: The term *generic* denotes a special usage of a referring expression, such that not a particular individual or object is meant, but rather a class of entities or features of this class.

   (10.a) *WhalesGEN are mammalsPRED.*

(10.b) *Der Pr¨asident wurde immer schon durch die Stimmenmehrheit bestimmt.*
"The PresidentGEN has always been elected by majority vote1,NEW."

Generics should not be annotated with a discourse referent index – unless they are subsequently referred to:

(10.a') *Whales1,GEN are mammalsPRED. They1,OLD descend from land animalsGEN.*

This includes both nominal and pronominal markables. Generic pronouns such as *we*, *you*, *they* (in cases where they do not carry a specific reference), *someone*, *anyone*, *one*. Cf. German *man*.

(11.a) *Meier said to Müller: "[You]GEN should go now."*

(11.b) *Meier sagte zu Müller: „[Man]GEN sollte jetzt gehen."* (German)

(11.c) *Meier said to Müller: "Last year, [they]GEN demolished a house here."*

(11.d) *Meier sagte zu Müller: „Letzes Jahr haben [sie]GEN hier ein Haus abgerissen."* (German)

8. EXPL: Non-referring expression: Expletive expressions (English *it*) and pronominal adverbs that are controllers of relative clauses

   (12.a) *It was raining.*

   (12.b) *[It] was also considered certain that . . .* (English)

   (12.c) *[Es] galt zudem als sicher, dass . . .* (German)

9. PRED: Non-referring expression: Predicative NPs in copular sentences

   (13.a) *Nicht, dass beide eine Mehrheit für ihre Koalition suchten, war [das Ärgerliche in den vergangenen Tagen] ...* (German)

   (13.b) *The chief physician was [a real professional]NM.*

   (13.c) *Max Müller is [the greatest center forward of all time]NM!*

10. IDIOM: Non-referring expression: apparent referring expressions (e.g., definite NPs) in fixed, conventionalized idioms and corresponding collocations:

    (14.a) *jemandem auf die Nerven gehen* (German, "to annoy someone")

    (14.b) *Er brachte mich auf [die Palme]NM* (German)

    (14.c) *Und dann warf sie [die Flinte]NM [ins Korn]NM.* (German)

    Note: Referring expressions in productive, transparent metaphors that are sufficiently transparent should be annotated like anaphoric expressions. The annotator may add AMBIG:IDIOMif not sure about their annotation. In (7.d), *der Spatz in der Hand*, a definite NP in German, can be generic, part of an idiom, or referring:

(14.d) *Lieber [der Spatz in der Hand] als [die Taube auf dem Dach]* (PCC, 12666) "A bird in the hand is worth two in the bush" (Context: a mayor finds an investor for his town willing to make only minimal investments).

(14.e) *So lässt sich [das schlingernde City-Schiff]PM vielleicht doch noch auf einen erfolgversprechenden Kurs bringen.* (German, maz-18914, here, a reference to a city is made, but combined with the metaphorical image of a ship in troubled water, for which the substitution test would fail)

11. `other`: other, non-referring expression, please provide a description in round parentheses. Includes, for example, NPs under the scope of a negation that cannot be referred to

    (15) *I didn't buy [a new car]NM after all.*

4.6 Example

(16.a) *[Die einstige Fußball-Weltmacht]1,CAT zittert [vor einem Winzling]2,CAT,AMBIG(2,6).* "[The former football World Power]d is shivering [in the face of a mite]s."

(16.b) *Mit seinem3,BOUND Tor4,NEW zum 1:05,NEW für die Ukraine6,NEW stürzte der 1,62 Meter große [Gennadi Subow]2,NEW die deutsche Nationalelf1,NEW vorübergehend in ein Trauma7,NEW.* "By [his]s goal that set the score to 1:0 [for Ukraine]u pitched [Gennadi Subow]s, 1.62 Meter tall, [the German National Eleven]d in a shock for a while…"

(16.c) *Je kleiner die Kicker2,OLD,AMBIG:COREF(2,1) daherkommen, desto größer wird der Gegner1,OLD,AMBIG:COREF(1,2) geredet...* (German, maz-10374) "The smaller the kickers appear, the greater [the rivals]d?/u? are rumoured to be." (PCC, 10374)

Note that here, the antecedent of die Kicker "kickers" depends on the understanding of the "size" metaphor, it can be either the Ukrainian team (presented as having short players), or the German team (which has not been favored in the first match), or a generic description (which would mean that the sentence is not directly linked with the discourse).

4.6 Trouble Shooting

4.6.1 Recurring Group Reference

Here is a very compact, constructed example:

(17.a) Peter1 and Malte2 went for a walk3. Both4>1,2 wore hats5.

(17.b) Peter1 had a coat6, Malte2 a rain jacket7.

(17.c) They**4** reached...

When annotating *They*, note that this group has been previously established. For this reason, we do *not* refer to the second respective mentions of *Peter* and *Malte*, but instead to the previously established index for the group introduced when annotating *Both*.

4.6.2 Quantified NPs

Quantified expressions are either OLD/GROUP or NEW:

1. They are OLD if the same group has been previously referred to.
2. Otherwise, they are GROUP if they describe a finte set of entities and all these entities are mentioned before individually.
3. Otherwise, they are OLD if they clearly delimit the 'set of objects'.
4. Otherwise, they are NEW.

Leaving the first two cases aside, a substitution substitution test helps with the decision about OLD and NEW: If we can insert a definite article or demonstrative pronoun, does that change the meaning? If not, this is referring expression.

(18) *people* –→ *all these people* → definite description –→ referential

4.6.3 Pronominal Adverbs

Depending on context, some words can be either referring expressions or not. This may be automatically pre-annotated, but must be marked as non-referring in their referentiality annotation (see above).

A notorious problem in German is the annotation of pronominal adverbs such as *damit* (in the sense "so that", not in the sense "with it"), if they act as a connector:

(19.a) *[Auf dem Tisch]PM liegt [eine Kneifzange]SM. [Damit]PM kann man viel anfangen.* (German, referential *damit* "with it")

(19.b) *[Ich]PM habe [dir]PM [den Brief]PM gezeigt, damit [du]PM bescheid weißt.* (German, non-referential *damit" "so that")

4.6.4 Relative Possessive Pronouns

Relative pronouns are syntactically bound and not to be annotated, but relative posessive pronouns in possessive constructions are treated as possessive pronouns.

Test for German: Ein (nicht-possessives) Relativpronomen ist genau dann gegeben, wenn es durch ‚*welch*' ersetzt werden kann:

(20.a) *Und so schielten [die Israelis] [nach Washington], [an [dessen]/∗welchem Tropf] [sie] hängen.* (maz-19074)

(20.b) *Und so schielten [die Israelis] [nach Washington, das/welches [sie] wirtschaftlich stützt].*

4.6.5 Cataphora

We distinguish two types of forward-referring expressions, discourse cataphora and syntactic cataphora.

4.6.5.1 Discourse Cataphora (Anaphora of Anticipation)

Discourse cataphora is a label used for non-pronominal reference forward. Sometimes an author introduces a discourse referent by means of an underspecified NP, i.e. an NP that cannot be interpreted only on the basis of the reader's knowledge up to this point. This way the author tries to encourage the reader to continue reading, in order to catch up the missing information. In the example below, *die einstige Fußball-Weltmacht* and *vor einem Winzling* should be annotated as discourse cataphors, since their referents cannot be identified until introduced explicitly in the following text (*Deutschland* and *Ukraine* correspondingly).

(21) *Die einstige Fußball-Weltmacht zittert vor einem Winzling* (newspaper article title)

In case one goes on reading the text, it becomes clear that *die einstige Fußball-Weltmacht* refers to Germany, whereas *ein Winzling* refers either to the Ukraine or the 1.62 meter tall ukranian footballer who made the most impact in the match [5]. Discourse cataphors have to be annotated as normal anaphors, i.e. in accordance with the Chain Principle (p. 12), i.e. the most recent referent mention to the left (if any) is considered to be an antecedent.

4.6.5.2 Syntactic cataphora

(22) *Through [his] lawyers, [Mr. Antar] has denied allegations in the SEC suit ...* (WSJ)

Syntactic cataphors are to be annotated like pronominal anaphora, mark referentiality as `BOUND` or `OLD`, whichever appropriate.

The following examples (a nominal head followed by a restrictive modifier), although traditionally classified as cataphora, should NOT be annotated as such.

(23) ... [*the car that went through his garden wall*]...
(24) ... [*the patterns of industrial development in the U.S*]....

In case of doubt between syntactic cataphora or anaphora, decision has to be made as follows.

(25.a) *Die einstige Fußball-Weltmacht zittert [vor einem Winzling]s.*

(25.b) *[Mit [seinem]s Tor zum 1:0 fu¨r die Ukraine] stu¨rzte [der 1,62 Meter große Gennadi Subow]s [die deutsche Nationalelf] voru¨bergehend in ein Trauma.*

In the example, *seinem* refers to *Gennadi Subow* who was introduced in the very first sentence as *vor einem Winzling*. Following the preferences, we establish an anaphoric (cataphoric) link to the right. Thus, the anaphoric chain looks as follows:

- *seinem* → *Gennadi Subow* (same-sentence)
- *Gennadi Subow* → *vor einem Winzling* (right+previous, Chain Principle)

5. Information Status

This document is a slightly revised version of the Coding Protocol for Statuses on the Givenness Hierarchy according to Gundel et al. (1993), revision of April 2023, see Readme for authors and contributors. Note that the criteria in this coding protocol are sufficient, not necessary conditions for assigning a particular status.

5.1 Givenness Hierarchy

The statuses of the Givenness Hierarchy (Gundel, Hedberg and Zacharski 1993) describe degrees of accessibility of discourse referents at a given point in discourse. In the literature, this phenomenon is referred to as *givenness*, *accessibility*, *salience*, etc. Throughout this manual, we use the term **information status** according to Lambrecht (1996). Gundel et al. originally used the term *cognitive status*.

Information status is a property of cognitive entities/mental representations. The terms IN FOCUS, ACTIVATED, FAMILIAR , UNIQUELY IDENTIFIABLE, REFERENTIAL, and TYPE IDENTIFIABLE each describe an information status on the Givenness Hierarchy (Gundel, Hedberg and Zacharski 1993). A referent IN FOCUS is considered to be higher on the hierarchy than a referent that is only TYPE IDENTIFIABLE, for example:

- IN FOCUS > ACTIVATED > FAMILIAR > UNIQUELY IDENTIFIABLE > REFERENTIAL > TYPE IDENTIFIABLE

When determining the information status using the protocol, imagine you are the speaker/writer and ask yourself what you can assume about the information status of the intended interpretation/referent for the addressee at the point just before the form is encountered. Annotate sentence by sentence. Check the criteria for each status in the order they are listed below:

1. Start with the status IN FOCUS.
2. If none of the criteria apply, try ACTIVATED.
3. If none of the criteria apply, try FAMILIAR.
4. If none of the criteria apply, try UNIQUELY IDENTIFIABLE.
5. If none of the criteria apply, try REFERENTIAL.
6. If none of the criteria apply, try TYPE IDENTIFIABLE.

Stop when you find a criterion that applies. This is the highest information status for the referent/interpretation you are checking.

After annotating all referents of the current sentence, annotate the CB (backward-looking center, familiarity topic) according to Centering Theory (Grosz et al. 1995).

5.2 IN FOCUS (IS=FOCUS)

A referent is IN FOCUS[1] if it meets at least one of the following criteria:

1. It is the interpretation of the main clause subject or the syntactic topic in the immediately preceding sentence/clause (syntactic topics include topicalized or dislocated phrases, including topic marked phrases, e.g. the *wa* phrase in Japanese).

   (1) Midge pushed thick, wiry black hair back from her square forehead with a sturdy brown arm. Nothing unsubstantial or fairylike about her. (From *Murder after Hours*, Agatha Christie)
   (2) John Kerry lost in Ohio. This cost the Senator the election.
2. It is part of the interpretation of a previous part of the same sentence.

   (3) You can wear my scarf if you can find it.
   (4) If you stand on this chair, the chair will break.
3. It is the interpretation of the syntactic focus of the immediately preceding clause (i.e., postcopular position of a cleft or existential sentence).

   (5) There was a mouse on the table. It was very large.
   (6) It was the dog that Bill was afraid of. He was very large.
4. It is a higher level topic that is part of the interpretation of the preceding clause (whether it is overtly mentioned there or not).

   (7) The kitchen has a new countertops and a beautiful tile floor. There's also a big walk-through closet. Would you like to take a look at it? Both the kitchen (criterion 4) and the closet (criterion 3) are in focus.
5. It is part of the interpretation of each of the two immediately preceding clauses.

   (8) It was the dog that Bill was afraid of. Small animals didn't usually frighten Bill. He was very large. (*him* most likely to be interpreted as Bill, not the dog)
   (9) A: She will be nice to Gerda and she will amuse Henry, and she'll keep John in a good temper and I'm sure she'll be most helpful with David – B: David Angkatell? A: Yes. He's just down from Oxford. (From *Murder after Hours*, Agatha Christie)
6. It is the event denoted by the immediately preceding sentence.

   (10) John fell off his bike. This/it happened yesterday.

5.2.2 ACTIVATED (ACTIVATED)

A referent is ACTIVATED if it meets one of the following criteria.

1. It is part of the interpretation of one of the immediately preceding two sentences.

   (11) Central to the case was a Lewinsky-Tripp conversation that Mrs. Tripp taped on Dec. 22, 1997. This was the last talk between the two women that Mrs. Tripp recorded.
2. It is something in the immediate spatio-temporal context that is activated by means of a simultaneous gesture or eye gaze.

   (12) (looking at the wrench) Please hand me that (wrench (over there))

3. It is a proposition, fact, or speech act associated with the eventuality (event or state) denoted by the immediately preceding sentence(s).

    (13) A. John fell off his bike. B. That's not true.
    (14) A. John fell off his bike. B. Can you say that again?

5.2.3 FAMILIAR (FAMILIAR)

A referent is FAMILIAR if it meets one of the following criteria.

1. It was mentioned at any time previously in the discourse.

    (15) A Phillipine Airlines jet with 290 people aboard was hijacked today by a man who took everyone's money and then parachuted to the ground outside Manila's airport and the passengers were let off safely. The jetliner left Davao City, in the southern Phillipines, for the 90-minute flight to Manila with 278 passengers and 12 crew aboard, PAL said. The hijacker, wearing a blue ski mask and carrying a handgun…
2. It can be assumed to be known by the hearer through cultural/encyclopedic knowledge or shared personal experience with the speaker.

    (16) If one takes a step back and looks at the rest of this week's music-group news, the situation looks bad for ugly, unpredictable rock 'n' roll: one of the most popular American rock bands of the 90's.

5.2.4 UNIQUELY IDENTIFIABLE (UNIQUE)

A referent is UNIQUELY IDENTIFIABLE if it meets one of the following criteria:

1. The referring form contains adequate descriptive/conceptual content to create a unique referent.

    (17) s: hello can I help you u: yeah I want t- I want to determine the maximum number of boxcars of oranges that I can get to Bath by 7 a.m. tomorrow morning so hm so I guess all the boxcars will have to go through oran- through Corning because that's where the orange juice factory is [Trains Corpus. Heeman & Allen 1995]
2. A unique referent can be created via a 'bridging inference' by association with an already activated referent.(e.g. A house….the front door)

    (18) She got into bed, laid her head on the pillow, and in two minutes was sleeping like a child. (From *Murder after Hours*, Agatha Christie)
    (19) (Looking at a box) I think the bottom fell out.

5.2.5 REFERENTIAL (REF)

A referent exists, is REFERENTIAL, if it meets one of the following criteria.

1. It is mentioned subsequently in the discourse.

(20) When my youngest child was 3 or so, we were at a friend's house visiting and my friend was babysitting her infant nephew.

2. It is evident from the context that the speaker intends to refer to some specific entity.

(21) I want to tell you about this strange guy I saw today.

5.2.6 TYPE IDENTIFIABLE (TYPE)

An interpretation is TYPE IDENTIFIABLE if the sense of the phrase (the descriptive/conceptual content it encodes) is understandable.

(22) I don't have a VCR and neither does my neighbor.
(23) Whenever Mary passes that store, she always picks up a newspaper.

6. Information Structure

We provide a partial annotation of information structure, only, by focusing on information status and familiarity topics. As for the latter, we adopt the approach and the terminology of Centering Theory, and thus speak of "backward-looking center" (CB)

6.1 Familiarity Topic: Backward-Looking Center (CB)

After IS annotation, CB annotation is to be done in the `CB` column.

In Centering Theory (Grosz et al. 1995), the "backward-looking center" is a technical term for the notion of "familiarity topic". The following criteria apply: 1

- Each sentence ("utterance") has at most one backward-looking center.
- The backward-looking center of the current sentence must be explicitly mentioned ("realized") in the immediately preceding sentence. That is, it must have been previously annotated as `IN FOCUS` or `ACTIVATED`.
- If there is more than one CB candidate that has been mentioned in the preceding sentence, check the properties of its antecendent. If an IN FOCUS expression refers to the preceding sentence or clauses within it, annotate it only as CB if no other candidate an be found.
- Mark the expression as CB whose antecedent is highest on the following ranking ("salience ranking"): 2
    1. SUBJECT (of main clause, `GR=SBJ`)
    2. OBJECT (of main clause, e.g., direct or indirect object, `GR=OBJ`)
    3. OTHER (oblique argument of main clause, e.g., prepositional phrase, `GR=other`),
    4. MAIN CLAUSE (event anaphora: refer to the main clause or the full sentence, rather than any of its arguments, no `GR` annotation)
    5. SUBJECT (of dependent clause, `GR=SBJ_2`, `SBJ_3`, etc.)
    6. OBJECT (of dependent clause, `GR=OBJ_2`, `OBJ_3`, etc.)
    7. OTHER (of dependent clause, `GR=other_2`, `other_3`, etc.)

8. DEPENDENT CLAUSE (event anaphora: refer to a dependent clause, no GR annotation)
9. etc., for more deeply embedded dependent clauses

- If there are multiple CB candidates whose antecedent realization (according to this ranking) is identical, chose the one whose antecedent is mentioned *first* in the preceding sentence.

CB annotation is partially pre-annotated, but has to be manually refined.

Selected CB examples: - antecedent SUBJECT (main clause, Grosz et al. 1995, ex. 6)

```
> (1.a) * **Susan**  gave Betsy a pet hamster.*

> (1.b) <ins>She</ins> reminded her that such hamsters were quite shy.
```

- antecedent direct object (main clause, Grosz et al. 1995, ex. 18)

  (2.a) *I'm reading **The French Lieutenant's Woman**.*

  (2.b) *The book, which is Fowles's best, was a bestseller last year.*

- antecedent indirect object (main clause, Grosz et al. 1995, ex. 17)

  (3.a) *My dog is getting quite obstreperous.*

  (3.b) *I took **him** to the vet the other day.*

  (3.c) *The mangy old beast always hates these visits.*

- antecedent clause

  (4.a) * **John fell off his bike**.*

  (4.b) *This/it happened yesterday.*

- antecedent SUBJECT (dependent clause, Grosz et al. 1995, ex. 2)

  (5.a) *It was a store **John** had frequented for many years.*

  (5.b) *He was excited that he could finally buy a piano.*

A. Supplemental

A.1 Notes

1. Background and Terminology

   – 1: In the original PoCoS/PCC guidelines, markables were defined as phrasal expressions. Here, we annotate syntactic heads, instead.

   – 2: The head-based annotation adopted in these guidelines is an innovation to facilitate interoperability with Universal Dependency annotations. Krasavina and Chiarcos (2007) and Chiarcos et al. (2016) focused on the annotation of phrases, instead.

2. File Format ands Editing

3. Automated Pre-Annotation of Markables

   - 1: The definition of primary markables follows Krasavina and Chiarcos (2007). Chiarcos et al. (2016) singled out non-referential markables from primary markables as they do not rely on automated pre-annotation.

   - 2: From UD annotation, we cannot extract times and dates reliably. So, these receive no special handling as primary markables (different from Stede et al. 2015).

   - 3: Chiarcos and Krasavina (2005) also include zero (pro-drop) pronouns under pronouns. Here, we follow token-based annotation, so that zeros should not be annotated.

     (43) *Johnj stepped in the kitchen, Øj opened the fridge and Øj decided NO-ZERO to take a pizza.*

     Here, that John is the (implicit) subject of the clause *to take a pizza*. However, this is not an instance of ∅-pronoun, since the insertion of *John* (no matter at which position within the phrase) would make the utterance ungrammatical. If not sure whether to annotate a ZERO or not, try to insert a full description of the corresponding referent. Note that zeros have to be sentential arguments, no adjuncts.

4. Nominal coreference

   - 1: Our annotation of anaphora as coreference differs from Chiarcos et al. (2016) who annotated anaphoric relations between anaphors and their antecedents, instead. Note that this means that we do not distinguish anaphoric (pronominal) and non-anaphoric (nominal) coreference, here.
   - 2: `REF=OLD` corresponds to "discourse old" according to Prince (1992). Originally abbreviated as `referring` (Chiarcos and Krasavina 2005).
   - 3: `REF=NEW` corresponds to "discourse new" according to Prince (1992). Originally abbreviated as `discourse-new` (Chiarcos and Krasavina 2005).
   - 4: `REF=CAT`, originally abbreviated as `discourse-cataphora` by Chiarcos and Krasavina (2005).
   - 5: Groups and situational references were originally subsumed under `other` in the PoCoS core scheme (Chiarcos and Krasavina 2005)
   - 6: Bound pronouns were part of the PoCoS extended scheme, not the core scheme (Chiarcos and Krasavina 2005)
   - 7: Abbreviated `ambig-ante` in Chiarcos and Krasavina (2005)

5. Information status

   - 1 In the context of information structure, the naming "in focus" or "focus" for the state with maximum givenness is very unfortunately. We stick with

Gundel et al.'s terminology, but keep in mind that we are talking about the focus of attention here, not focus in the information-structural sense.

6. Topic

   – 1: Following Grosz et al. (1995), Centering Theory was extended and parameterized. The definitions given above represent *one* specific interpretation of Grosz et al.'s criteria designed to facilitate unambiguous annotation. Other interpretations are possible.
   – 2: The salience ranking has been extended to include dependent clauses and word order, following Gernsbacher (2013). Furthermore, the Centering category "OTHER" is split into "OBLIQUE ARGUMENT" and "CLAUSE". Unlike the original formulation of Centering, the Givenness Hierarchy supports event anaphora.

A.2 Sources

The current manual has been compiled by Christian Chiarcos, University of Augsburg, in spring 2023. See accompanying readme for authors, contributors and revision history.

Sections 3 and 4 and parts of Sect. 1 are based on

- Christian Chiarcos and Olga Krasavina (2005), PoCoS. Potsdam Coreference Scheme, Tech. Rep., University of Potsdam, Germany
- Olga Krasavina and Christian Chiarcos (2007), PoCoS. Potsdam Coreference Scheme, First Linguistic Annotation Workshop (LAW-2007), held in conjunction with ACL-2007, Prague, Czech Republic, June 2007
- Christian Chiarcos, Manfred Stede, Saskia Warzecha (2019), Nominale referentielle Ausdrücke, In: Stede, M. (Ed.). (2016). Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0 (Vol. 8). Universitätsverlag Potsdam, p.55-70
- Christian Chiarcos, Manfred Stede, Saskia Warzecha (2019), Nominale Koreferenz, In: Stede, M. (Ed.). (2016). Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0 (Vol. 8). Universitätsverlag Potsdam, p.71-85

Whenever we draw from these texts, this is not specifically marked. These texts and the current manual represent different developmental stages and instantiations of the PoCoS core scheme (Chiarcos and Krasavina 2005),

Section 5 is based on Gundel et al.'s Coding Protocol for Statuses on the Givenness Hierarchy (2006, http://www.sfu.ca/~hedberg/Coding_for_Cognitive_Status.pdf, accessed 2023-04-16). Except for editorial updates, this document is largely unchanged. Changes are not explicitly marked, but are documented in Git history.

Sections 2 and 6 have been written from scratch for this manual by Christian Chiarcos, Spring 2023.

A.3 Literature References (Selection)

Biber, Douglas et al.. *Longman Grammar of Spoken and Written English*. Longman, 1999.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. Computational linguistics, 47(2), 255-308.

Gernsbacher, M. A. (2013). *Language comprehension as structure building*. Psychology Press.

Gibbs, R.W. *The poetics of mind*. Cambridge University, Cambridge, 1994. Gardent, Claire, H ´el`ene Manu´elian, and Eric Kow. Which bridges for bridging descriptions. In *EACL Workshop on Linguistically Interpreted Corpora proceedings.*, 2003.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, *21*, 203-225.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274-307.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge university press.

Mitkov, R. et al.. Coreference and anaphora: Developing annotating tools, annotating resources and annotations strategies. In *Proc. DAARC 2000*, pages 49–58, Lancaster, UK, 2000.

Reinhart, T. (1980). Conditions for Text Coherence. Poetics Today, 1(4), 161–180. https://doi.org/10.2307/1771893

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual. Philadelphia, University of Pennsylvania, 35, 108.

About this Document

The Augsburg Manual for the Annotation of Reference and Information Structure (AURIS) provides a practically-oriented guideline for the annotation of reference and selected aspects of information structure.

At the moment, it comprises guidelines for

- the automated annotation of referring expressions,
- the manual annotation of coreference,
- the manual annotation of information status (“givenness”), and
- the manual annotation of familiarity topics (“backward-looking centers”).

The extension to other aspects of information structure is foreseen, but has not been implemented so far.

Note that all annotations need to revised if changes to this manual occur. For this reason, annotators must **NEVER** change this document. Instead, if you encounter difficulties or problems with the annotation, document your requirements or issues in an accompanying

document, along with a reference to the data where you encountered the issue and a brief statement on how you solved or marked it.

Content

The manual consists of five separate documents:

- terms.md: basic terminology, originally by Chiarcos and Krasavina (2005)
- format.md: file format and annotation procedure, by Christian Chiarcos (2023)
- refexp.mp: guidelines for automated pre-annotation for referring expressions, originally by Chiarcos and Krasavina (2005)
- coref.md: guidelines manual annotation of coreference and referentiality, originally by Chiarcos and Krasavina (2005)
- information-status.md: guidelines for the manual annotation of information status ("givenness"), originally by Gundel, Hedberg and Zacharski (1993)
- topic.md: guidelines for the annotation of the familiarity topic according to Centering Theory (backward-lookinf center, Grosz et al., 1995)

Supplementary material is provided in

- lit.md: Sources, references, footnotes
- addenda.md: Material for future extensions

This document is meant to be a practical handbook, compiled and revised from earlier manuals, with a focus on examples and common problems. In some design decisions, we deviate from our sources:

- **referring expressions**: We skip the coreference annotation principles in order to provide a more minimal description. We introduced head-based annotation instead of phrase-based annotation. We skip the extended part of PoCoS. We extend the annotation to antecedents of event anaphora.
- **coreference**: We skip the coreference annotation principles in order to provide a more minimal description. Instead of anaphoric relations, we annotate coreference by coindexing. We skip the extended part of PoCoS. We skip features of referring expressions than can be derived from syntax.
- **information status**: We introduced head-based annotation. The original hierarchy is extended for backward-looking center.

Original contributions include specifics of format and annotation procedure.

Disclaimer

The AURIS guidelines have been compiled from existing guidelines whose original creators are attributed along with the compilator. Note that we do not track all possible changes within this document, but instead, we provide the modified guidelines along with the original files in the same format, so that changes can be tracked automatically (e.g., using the command line tool `diff`).

Because they can be automatically identified, we do not mark literal quotations from the original document. Note that this manual is **not to be published** unless this information is to be added to it.

Contributors

Authors

We list direct contributors as well as primary authors of sources that went, fully or partially, into this document, along with their (estimated) duration of involvement

- CC: Christian Chiarcos (coreference/English, German: since 2005, information status: since 2008), University of Augsburg, Germany
- OK: Olga Krasavina (coreference/English, Russian: 2005-2011), HU Berlin, Germany / Moscow State University, Russia
- MS: Manfred Stede (coreference/German: 2005-2015), U Potsdam, Germany
- SW: Saskia Warzecha (coreference/German: until 2015), U Potsdam, Germany
- JG: Jeanette Gundel (information status: 1993-2006)
- NH: Nancy Hedberg (information status: 1993-2006)
- RZ: Ron Zacharski (information status: 1993-2006)

Other Contributors

Other contributors are people who provided feedback and input, incl. annotators, in alphabetical order:

- Mamadou Bassene (information status: before 2007)
- Tonya Custis (information status: before 2007)
- Bryan Gordon (information status: before 2007)
- André Herzog (coreference/German: before 2016), U Potsdam. Germany
- Linda Humnick (information status: before 2007)
- David Kaupat (coreference/German: before 2016), U Potsdam. Germany
- Amel Khalfoui (information status: before 2007)
- Sara Mamprin (coreference/German: before 2016), U Potsdam. Germany
- Ann Mulkern (information status: before 2007)
- Bonnie Swierzbin (information status: before 2007)
- Shana Watters (information status: before 2007)
- Dmitry Zalmanov (coreference/Russian: 2005-2007), Moscow State University, Russia

Unfortunately, we do not have records of all contributors or annotators whose feedback and experiences went into this manual, directly or indirectly, especially for the years before 2016. If you find yourself to be missing, please drop us a line.

History of this Document

The original annotation guidelines for coreference were drafted in 2004 by Christian Chiarcos and Olga Krasavina for the annotation of the Potsdam Commentary Corpus of German newspaper commentaries (PCC) (Stede, 2004) and the RST Discourse Treebank of Wall Street Journal articles (WSJ) (Carlson et al., 2003).

After a series of annotation experiments, the PoCoS Core Scheme was applied to the PCC by two instructed annotators, students of linguistics, whose portions had an overlap of 19 texts (11%). Part of the WSJ corpus has been annotated in co-operation with A.A. Kibrik, Moscow State University, with fourteen instructed annotators, also students of linguistics. In addition, experimental annotations for Russian were created by Olga Krasavina and Dmitry Zalmanov.

The PoCoS coreference scheme was published as a technical report by Chiarcos and Krasavina (2005), and in a subsequent conference paper (Krasavina and Chiarcos 2007). The PoCoS coreference scheme defines an underspecified core scheme and optional extensions (originally for English and German, and experimental extensions for Russian, the current guidelines represent yet another, multilingual extension). The German instantiation of these guidelines was subsequently revised as part of the Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0 (Stede et al., 2016), and its updates have been partially included in this document.

The Augsburg Information Structure corpus adopts a different technical infrastructure, using tabular formats, spreadsheet-based annotation with automated pre-annotation and head-based annotation. As a result, the annotation guidelines were massively restructured by Christian Chiarcos in spring 2023, incorporating and revising all earlier versions available to us at the time.

For this version of the manual, we extended the guidelines by including the guidelines for the annotation of information status ("givenness") by Gundel et al. (1993) and guidelines for the annotation of the backward-looking center ("sentence topic"), following a rigid interpretation of Grosz et al. (1995). These parts of the schema were initially developed independently, and only integrated in the process of compiling this manual.

- terms.md: definitions
  - 2023-05-15: integrated Chiarcos and Krasavina (2005)
  - 2023-05-05: integrated Krasavina and Chiarcos (2007)
  - 2023-04-20: extracted from refexp.md and coreference.md, see there for sources and contributors [CC]
- format.md: file formats and annotation proceduce
  - 2023-05-05: first draft [CC]
- refexp.md: guidelines for automated pre-annotation
  - 2023-05-15: integrated Chiarcos and Krasavina (2005)
  - 2023-05-05: integrated Krasavina and Chiarcos (2007)
  - 2023-04-20: restructured and fully revised [CC]

    - 2015-xx-xx: translation of relevant parts of Stede et al. (2015) to English
        - source: Stede, M. (Ed.). (2016). Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0 (Vol. 8). Universitätsverlag Potsdam, p.55-88
        - authors (of chapters): Christian Chiarcos, Manfred Stede, Saskia Warzecha [CC, MS, SW]
        - contributors (of chapters): André Herzog, David Kaupat and Sara Mamprin
        - excerpt and (minor) revision of the Potsdam Coreference Scheme (PoCoS), limited to German
    - 2007-06-28: Krasavina, O., & Chiarcos, C. (2007, June). PoCoS-Potsdam coreference scheme. In Proceedings of the Linguistic Annotation Workshop (LAW-2007), held in conjunction with ACL-2007, Prague, Czech Republic (pp. 156-163). [OK,CC]
        - Reference publication for the Potsdam Coreference Scheme (PoCoS)
    - 2005-10-25: Chiarcos, C., & Krasavina, O. (2005). Annotation Guidelines PoCoS–Potsdam Coreference Scheme. Technical Report. University of Potsdam, Germany. [CC,OK]
        - Original edition of the Potsdam Coreference Scheme (PoCoS), focusing on English, Russian and German
- coreference.md
    - 2023-05-15: integrated Chiarcos and Krasavina (2005)
    - 2023-05-05: integrated Krasavina and Chiarcos (2007)
    - 2023-04-20: revision by Christian Chiarcos, fully restructured
        - The original annotation procedure was focusing on anaphoric relations. Now revised to annotate coreference sets. However, many examples are taken over.
    - 2015-xx-xx: translation of relevant parts of Stede et al. (2015) to English
        - see under refexp.md (above) for sources and contributors
- information-status.md
    - 2023-05-16: moved backward-looking center into separate document: topic.md
    - 2023-05-05: backward-looking center as a separate layer of annotation
    - 2023-04-16: revision by Christian Chiarcos, extended for backward-looking center
    - 2006-05-xx: revision of May 2006
        - published under http://www.sfu.ca/~hedberg/Coding_for_Cognitive_Status.pdf
        - authors: Jeanette Gundel, Nancy Hedberg, Ron Zacharski
        - contributors: Ann Mulkern, Tonya Custis, Bonnie Swierzbin, Amel Khalfoui, Linda Humnick, Bryan Gordon, Mamadou Bassene, Shana Watters
    - 2004-07-xx: preceding revision of Jeanette Gundel, Nancy Hedberg, Ron Zacharski of July 2004

  - 1993: original version by Gundel, Hedberg and Zacharski
- lit.md: References
  - currently from Chiarcos and Krasavina (2005), only
  - 2023-05-15: initial version, from Chiarcos and Krasavina (2005)
- problems.md: Problematic cases
  - currently from Chiarcos and Krasavina (2005), only (needs to be completely re-analyzed)
  - 2023-05-15: initial version, from Chiarcos and Krasavina (2005)
- topic.md
  - 2023-05-16: extracted from information-status.md (CC)