ARRAU 3 Annotation Manual
Version 1.0

Massimo Poesio, Maris Camilleri, Paloma Carretero-Garcia and Ron Artstein
Revised for ARRAU3 from the original manual by Ron Arstein and Massimo Poesio,
with additional input from Antonella Bristot, Federica Cavicchio, Francesca Delogu,
Kepa Rodriguez and Olga Uryupina

November 8, 2021

A general introduction to the ARRAU annotation

The goal of the ARRAU annotation[1] is to collect data about the way context is constructed and used in natural language, particularly in relation to two phenomena normally called (nominal) ***anaphoric*** and ***deictic reference***.

The primary use of the term 'anaphora' is to describe what might also be called ***entity context dependence***. Many expressions of natural language are context-dependent, in the sense they derive at least part of their meaning from their context. Anaphoric expressions are so-called because their interpretation involves relations to the ***(discourse) entities*** contained in the context, whether implicitly or explicitly. A clear example of expressions that depend for their interpretation on the entities in the context are ***pronouns***. The meaning of a pronoun is typically (although not always) related to an entity already introduced in the linguistic context (e.g., as the result of a fuller linguistic expression produced as part of the previous utterances in the conversation). For example, in the fragment below, the pronoun *it* (underlined) refers to an entity already mentioned earlier in the same utterance (and introduced in the linguistic context) using the phrase *a boxcar* (in bold). This entity is called the ***antecedent*** of the pronoun.

> M: I want you to take [*a boxcar*] from [Elmira] and load
> [<u>it</u>] with [oranges]

As most anaphoric expressions have a reduced form in comparison with the forms used to introduce entities in context, by using them language users achieve coherence (they connect the current utterance to previous ones) in a parsimonious fashion.

[1] HISTORICAL NOTE: The last complete version of the ARRAU Annotation Manual was produced in June 2006 for ARRAU Release 1. Release 2 of ARRAU was annotated with reference to that manual, integrated with instructions from the final (unpublished) version of the GNOME annotation manual for the annotation of bridging, genericity, semantic category, and morpho-syntactic properties. This new version of the manual integrates the instructions from these different sources, and expands the instructions for bridging annotation, category, genericity, discourse deixis, and ambiguity, addressing a number of issues that emerged during the ARRAU 2 annotation. The scheme remains the same.

However, as said above, our primary interest in this study is not the (lexical) semantics of context-dependent expressions, but the way in which context is constructed and modified (an aspect of language often considered part of pragmatics). Therefore we are not only interested in cases in which the subsequent mention of an object is done using reduced forms; we want to track ***all*** mentions of an object through a text, including also cases in which these mentions are achieved using forms that normally would not be considered anaphoric, such as proper names like *Corning* or indefinite NPs like *a boxcar*.

(Anaphoric) context dependence can be more complex than in the example just seen. Some expressions derive their meaning from the linguistic context not in that they refer to an object already mentioned, but because they refer to entities which, while new in the discourse, are nevertheless tightly related to objects previously mentioned, in such a way that their reference would not be understandable if such relation wasn't apparent. An example is shown below: *the wheel* refers to an object which hasn't been mentioned before in the dialogue, but it is nevertheless understandable because it is a part of an object – *the boxcar at Elmira* – mentioned in the previous sentence.

```
S: Bad news about [the boxcar at Elmira]. [They] tell me
[the wheel] is broken and will have to be fixed
```

These expressions are usually called ***bridging*** references as the listener is required to 'bridge' the gap by identifying which relation holds between the expression and which previous expression. You will realize that a lot of expressions could be considered anaphoric in this sense; we only want to mark a few such cases however. These bridging references are not so common in the dialogues, but we will see more examples of them in the texts we will annotate later.

Language expressions may also depend on the visual, as opposed to verbal, context. An utterance like *Could you pass me* [*the salt*]*?*, uttered, e.g., at a restaurant, is usually understandable even if salt hasn't been mentioned before; the recipient can recover the referent of the expression *the salt* from the visual context. (An even clearer example is *Could you please close* [*the window*]*?*.) We will use the term ***deictic references*** to refer to these expressions. In the TRAINS dialogues in particular, but also in the GNOME-MUSEUM documents, conversational participants / the writer often refer to objects in the map using deictic references. We want to mark these references, both because it's an easy way to check the consistency of the annotation, and to get data about context-dependent expressions whose interpretation can only be recovered from the visual context.

In addition to anaphora and deictic reference, you will be asked to annotate a few types of information that are useful for understanding your judgments about anaphora and dexis.

Types of documents

The ARRAU corpus is a collection of documents of different genres, including:

- News articles from the Wall Street Journal (RST).

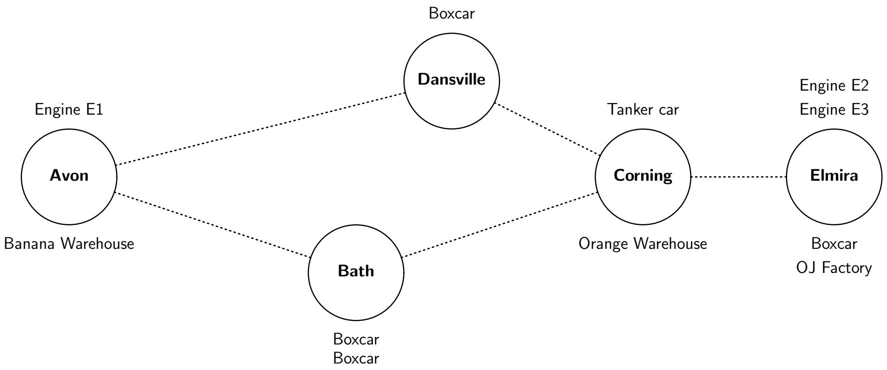


Figure 1: The TRAINS world

- Spoken dialogues from the TRAINS corpus. These are dialogues between two people who are trying to devise a plan for moving around rail cars in the railway network described in the map in Figure 1. You should consult the map as you read the dialogue.

- Transcripts of spoken narratives from the PEAR corpus. The narrator is describing a video in which a man climbs a pear tree with a ladder which is however taken away from a boy.

- Texts from catalogues of museum exhibitions from the GNOME-MUSEUM corpus. These include both descriptions of museum objects and more general texts providing background about artists or a particular artistic movement.

- Descriptions of medicines from the GNOME-PHARMA corpus. These are the leaflets that accompany medicines sold in the UK.

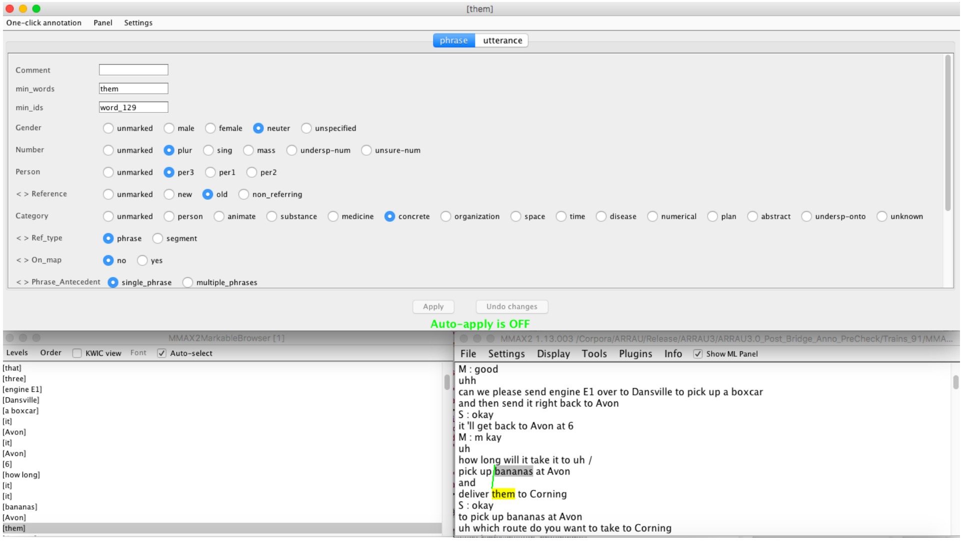
Although the core annotation tasks are the same for all genres, some of the attributes (and related instructions) only apply to some genres, as detailed below.

Procedure to follow during the annotation

§1 In this annotation, your task is to mark the ***anaphoric relations*** expressed by the ***markables*** in the documents you see on the screen, as well as several syntactic and semantic properties relevant to the study of these anaphoric relations. These markables include all the ***noun phrases*** in the document, as well as a few other constituents as explained below. To do this, you'll use the MMAX2 software, following the instructions below.

§2 Be sure to pay attention to the text while marking a document.

§3 This manual contains both instructions on how to do the annotation (with examples) and technical instructions for using the software.



Comment
min_words: them
min_ids: word_129
Gender: ○ unmarked ○ male ○ female ● neuter ○ unspecified
Number: ○ unmarked ● plur ○ sing ○ mass ○ undersp-num ○ unsure-num
Person: ○ unmarked ● per3 ○ per1 ○ per2
< > Reference: ○ unmarked ○ new ● old ○ non_referring
Category: ○ unmarked ○ person ○ animate ○ substance ○ medicine ● concrete ○ organization ○ space ○ time ○ disease ○ numerical ○ plan ○ abstract ○ undersp-onto ○ unknown
< > Ref_type: ● phrase ○ segment
< > On_map: ● no ○ yes
< > Phrase_Antecedent: ● single_phrase ○ multiple_phrases

□ KWIC view ☑ Auto-select

[that]
[three]
[engine E1]
[Dansville]
[a boxcar]
[it]
[Avon]
[it]
[Avon]
[6]
[how long]
[it]
[it]
[bananas]
[Avon]
[them]

☑ Show ML Panel

M : good
uhh
can we please send engine E1 over to Dansville to pick up a boxcar
and then send it right back to Avon
S : okay
it 'll get back to Avon at 6
M : m kay
uh
how long will it take it to uh /
pick up bananas at Avon
and
deliver them to Corning
S : okay
to pick up bananas at Avon
uh which route do you want to take to Corning

Figure 2: MMAX2 windows during annotation

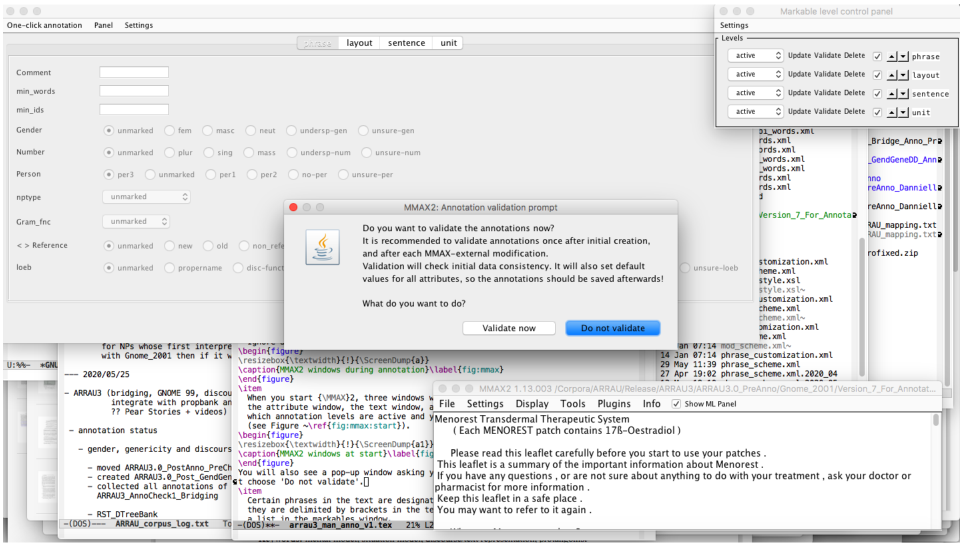
The MMAX2 software

The annotation tool you'll use is called MMAX2.

§4 For the annotation you will use three windows (Figure 2): the text window displaying the text in the document, the markables browser displaying all the markables, and the attribute window. Ignore all other windows that appear on your screen.

§5 When you start MMAX2, three windows will initially appear: the attribute window, the text window, and the level window which specifies which annotation levels are active and you should ignore (see Figure 3). You will also see a pop-up window asking you if you want to validate the document: just click on 'Do not validate'. You should then click on the 'Tools' menu in the text window, and select the 'Browsers' option, and then 'Markable Browser' to open up a new window, the markable browser (see Figure 4). After you've opened the markable browser, you should click on the 'Auto Select' button and choose the 'Document' option in the 'Order' menu so that the markables are presented in the order in which they appear in the document, although the alphabetical order is also useful in some circumstances (see Figure 5). At this point, you should be in the situation shown in Figure 2 and ready to start.

§6 Certain phrases in the text are designated as ***markables***: they are delimited by brackets in the text window, and also appear as a list in the markables window. Most markables you will be asked to annotate are noun phrases (NPs), with a few exceptions. ***You select a markable by clicking on it with the left mouse button in the markables window***; the corresponding phrase is then highlighted in the text window. You should go through the markables one at a time; the easiest way to do this is to simply enter 'Return' from the Markable Browser window, and


One-click annotation
Panel
Settings
phrase
layout
sentence
unit
Comment
min_words
min_ids
Gender
● unmarked ○ fem ○ masc ○ neut ○ undersp-gen ○ unsure-gen
Number
● unmarked ○ plur ○ sing ○ mass ○ undersp-num ○ unsure-num
Person
● per3 ○ unmarked ○ per1 ○ per2 ○ no-per ○ unsure-per
nptype
unmarked
Gram_fnc
unmarked
< > Reference
● unmarked ○ new ○ old ○ non_refe
loeb
● unmarked ○ propername ○ disc-funct
○ unsure-loeb
Markable level control panel
Settings
Levels
active
Update Validate Delete
☑ phrase
☑ layout
☑ sentence
☑ unit
MMAX2: Annotation validation prompt
Validate now
Do not validate
MMAX2 1.13.003 /Corpora/ARRAU/Release/ARRAU3/ARRAU3.0_PreAnno/Gnome_2001/Version_7_For_Annotat...
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel

Do you want to validate the annotations now?
It is recommended to validate annotations once after initial creation,
and after each MMAX-external modification.
Validation will check initial data consistency. It will also set default
values for all attributes, so the annotations should be saved afterwards!

What do you want to do?

```
    for NPs whose first interpre
    with Gnome_2001 then if it w

--- 2020/05/25

- ARRAU3 (bridging, GNOME 99, discou
      integrate with propbank an
      ?? Pear Stories + videos)

 - annotation status

   - gender, genericity and discours

     - moved ARRAU3.0_PostAnno_PreCh
     - created ARRAU3.0_Post_GendGen
     - collected all annotations of
       ARRAU3_AnnoCheck1_Bridging

     - RST_DTreeBank
-(DOS)---  ARRAU_corpus_log.txt   To
```

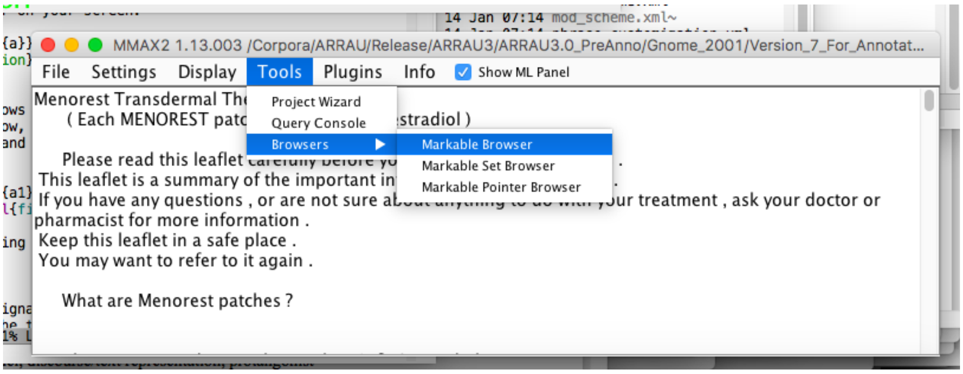
```
\begin{figure}
\resizebox{\textwidth}{!}{\ScreenDump{a}}
\caption{MMAX2 windows during annotation}\label{fig:mmax}
\end{figure}
\item
  When you start {\MMAX}2, three windows w
  the attribute window, the text window, a
  which annotation levels are active and y
  (see Figure ~\ref{fig:mmax:start}).
\begin{figure}
\resizebox{\textwidth}{!}{\ScreenDump{a1}}
\caption{MMAX2 windows at start}\label{fig
\end{figure}
You will also see a pop-up window asking y
t choose 'Do not validate'.
\item
  Certain phrases in the text are designat
  they are delimited by brackets in the te
  a list in the markables window.
-(DOS)**-  arrau3_man_anno_v1.tex   21% L2
```

```
01_words.xml        _Bridge_Anno_Pr
rds.xml             _GendGeneDD_Ann
ords.xml
_words.xml          nno
_words.xml          reAnno_Danniell
rds.xml
rds.xml             reAnno_Danniell
d
Version_7_For_Annota    RAU_mapping.txt
                        RAU_mapping.txt
                        rofixed.zip
stomization.xml
heme.xml
style.xsl
style.xsl~
ustomization.xml
cheme.xml
cheme.xml~
omization.xml
me.xml
an 07:14 mod_scheme.xml~
14 Jan 07:14 phrase_customization.xml
29 May 11:39 phrase_scheme.xml
27 Apr 19:02 phrase_scheme.xml.2020_04
```

Menorest Transdermal Therapeutic System
(Each MENOREST patch contains 17ß-Oestradiol)

Please read this leaflet carefully before you start to use your patches .
This leaflet is a summary of the important information about Menorest .
If you have any questions , or are not sure about anything to do with your treatment , ask your doctor or pharmacist for more information .
Keep this leaflet in a safe place .
You may want to refer to it again .

Figure 3: MMAX2 windows at start

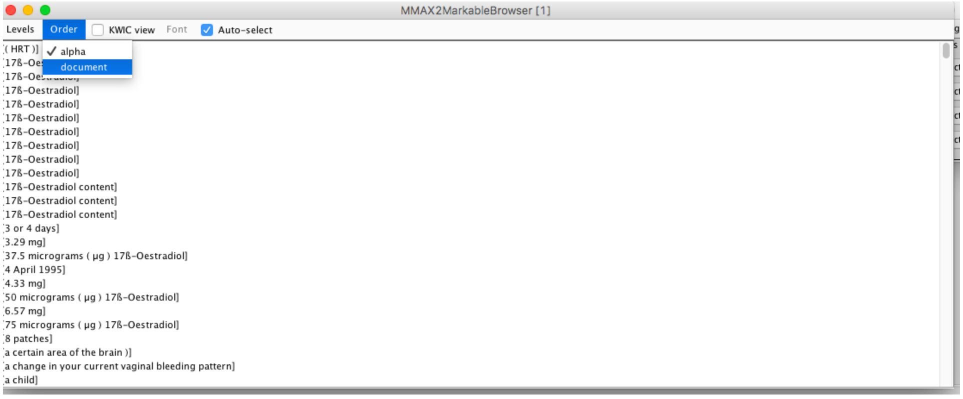


Menorest Transdermal The
(Each MENOREST patc stradiol)

Please read this leaflet
This leaflet is a summary of the important in
If you have any questions , or are not sure ab your treatment , ask your doctor or
pharmacist for more information .
Keep this leaflet in a safe place .
You may want to refer to it again .

What are Menorest patches ?

Figure 4: Opening the Markable Browser



(HRT)]
17ß-Oe
17ß-Oe
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol]
17ß-Oestradiol content]
17ß-Oestradiol content]
17ß-Oestradiol content]
3 or 4 days]
3.29 mg]
37.5 micrograms (µg) 17ß-Oestradiol]
4 April 1995]
4.33 mg]
50 micrograms (µg) 17ß-Oestradiol]
6.57 mg]
75 micrograms (µg) 17ß-Oestradiol]
8 patches]
a certain area of the brain)]
a change in your current vaginal bleeding pattern]
a child]

Figure 5: Choosing AutoSelect and markable order in the Markable Browser

you'll move on to the next markable. If you want to go back to a previous markable, scroll the window to the right place and then click left. (***NB: this is the behaviour if you clicked the 'Auto Select' button. If you didn't, you'll have to click on each markable twice to select it.***) Make sure you have selected the right markable before annotating its attributes as discussed below.

§7 Each markable is associated with a series of ***attributes***, and each attribute with a set of ***values***. When a markable is selected, the corresponding attributes appear in the attribute window. Initially, the values of all attributes are set to "unmarked".

§8 You should read the document, and as you do so go through the markables in order and mark the appropriate attributes, as explained below.

§9 The tool you are using has many features which are not described in this manual. If you do something accidentally and don't know how to get back, ask for help.

§10 You must select a value for each attribute that appears in the attribute window. To do this, you first have to select the markable by clicking on it in the markables window with the left mouse button, and then click on the appropriate radio button in the attribute window with the left mouse button.

Markable definition and identification

§11 We consider as markables all text constituents that could be interpreted as expressing a nominal anaphoric reference, or introducing a discourse entity that could serve as antecedent of an anaphoric reference. This includes:

- All noun phrases, including vacuous noun phrases, non-referring noun phrases, coordinations, etc.
- Appositive noun phrases, such as *the UK capital* in *London, [the UK capital] ...*
- All nominal modifiers that either refer directly, such as proper names (e.g., *UK* in *the [UK] parliament*), or are involved in anaphoric relations, as *stock market* in the following example:

  > .. small investors seem to be adapting to greater [*stock market*] volatility . . . Glenn Britta . . . is ''factoring'' [the market's] volatility ''into investment decisions.'' (*Amir Zeldes example from the ONTONOTES corpus*)

- Abandoned NP productions in conversations (see 'Incomplete' noun phrases below).

Excluded are those cases in which the 'anaphoric' expression is a subconstituent of its 'antecedent' whose interpretation is entirely recovered on syntactic grounds. This includes, in particular:

- Relative pronouns in restrictive relatives, such as *who* in *the man who shot Liberty Valance*.

§12 Markables will have been pre-identified for all documents, but the results of this pre-identification step are nor always accurate, even in cases such as the RST domain where markable identification was based on a manually annotated syntactic structure. In cases when an automatic mention detector was used, as in the case of the TRAINS documents, errors will be even more frequent. Markables identification errors include incorrect boundaries, spurious markables, and missing markables. One of your tasks is to correct such errors as required.

The attribute window

As you click on a markable that you haven't annotated yet, you will see in the attribute window a few attributes that are common to all markables. These include ***Comment***, ***min_ words*** and ***min_ ids***, a few attributes specifying the morphosyntactic properties of the noun phrase such as Gender, and ***Reference***. We will discuss each of these attributes next. *Depending on the values you specify for these attributes, and in particular for the **Reference** attribute, additional attributes may appear in the Attribute window; you should provide values for these as well.*

Comment

§13 You should use the ***Comment*** attribute to indicate when you are unsure about the annotation of a markable, and possibly what you think the problem was and / or which alternatives you considered.

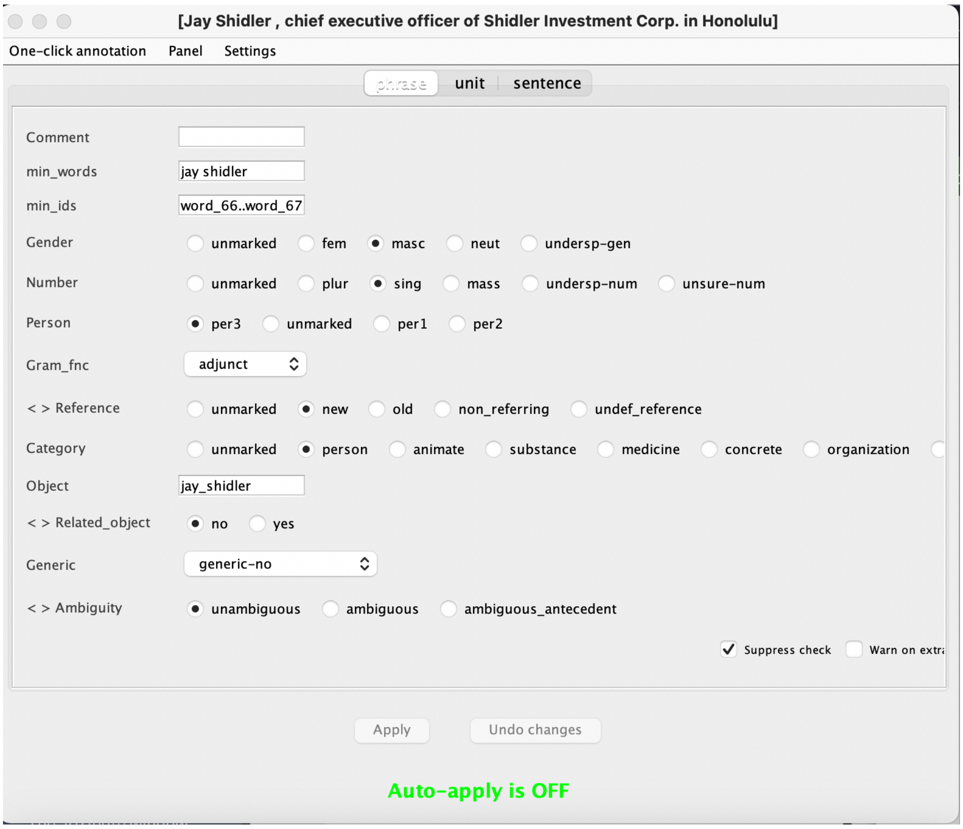
§14 Note that the auto-apply feature does not work for the comment attribute. After entering your comment, you must click on the "apply" button in order to save it.

Min information

§15 The attributes ***min_ words*** and ***min_ ids*** are used to specify the head of the markable (noun phrase). These fields are automatically pre-computed; your task is to check that values have been provided, and to let us know for which markable they are not, if any.

Exceptions and difficulties

- Please note that we follow an NP analysis rather than a DP analysis of noun phrases–i.e., we treat the noun as head instead of the determiner.



Comment

min_words jay shidler

min_ids word_66..word_67

Gender ○ unmarked ○ fem ● masc ○ neut ○ undersp-gen

Number ○ unmarked ○ plur ● sing ○ mass ○ undersp-num ○ unsure-num

Person ● per3 ○ unmarked ○ per1 ○ per2

Gram_fnc adjunct

< > Reference ○ unmarked ● new ○ old ○ non_referring ○ undef_reference

Category ○ unmarked ● person ○ animate ○ substance ○ medicine ○ concrete ○ organization

Object jay_shidler

< > Related_object ● no ○ yes

Generic generic-no

< > Ambiguity ● unambiguous ○ ambiguous ○ ambiguous_antecedent

☑ Suppress check □ Warn on extr

Figure 6: The ***min_ words*** attribute in the case of a markable with a multi-word head

- We assign a head to all noun phrases, with one exception: coordinated NPs such as *John and Mary*, which are generally considered not to have a head, and for which we do not mark Min information.

- Please note that the head is not always a single word. With proper names, the head is the entire proper name, which can consist of multiple tokens, as in *New York*, or *Philip Noiret*, or *Queen Mary University*. In these cases, you should check that the whole span is marked as done in Figure 6.

- Numerical determiners such as *one* or *two* can be used elliptically, as in *Chris found* [*two shells*] *and Pat only found* [*one*]. In such cases, you should mark the numerical determiner as head (***min_ words***) of the markable.

- Certain NPs have complex determiners such as *at least three* or *between five and ten*, but you should make sure the head noun is marked. E.g., in [*Between five and ten energy companies*] *will close this winter*, we consider *companies* the head of the markable.

Morphological information

Three attributes specifying morphological properties of the noun phrase are annotated: ***Gender***, ***Number***, and ***Person***.

§16 The attribute ***Gender*** specifies the (syntactic) gender of a markable. The possible values are "masculine", "feminine", "neuter", or "undersp-gender". This last value should be used when the NP could be either masculine of feminine (as in *the doctor*), with first and second person pronouns (*I*, *we*), with coordinated NPs containing both masculine and feminine NPs (as in *John and Mary*), or with organizations used metonymically to refer to the individuals working in them (*IBM announced they ....*). The test to be used to decide on the value for ***Gender*** is whether a subsequent pronoun co-referring with the NP would be masculine, feminine, neuter, or if more than one value could be used.

Tricky cases

- Because only singular pronouns are marked with gender in English, the pronominalization test suggested above doesn't work with plurals. In order to decide the gender of a plural NP, consider which pronoun would be used if the head noun were singular instead.
- In the case of NPs with coordinated plural heads (*any man or woman*), use the singular test to decide on the gender of each conjoined noun, then decide on the gender for the NP as a whole.

§17 The attribute ***Number*** specifies the (syntactic) number of a markable. We are mainly interested in the role of number in anaphora; as in the case of ***Gender***, therefore, the test to be used to decide on the value for ***Number*** is whether a subsequent anaphoric reference to the entity introduced by the NP would be singular or plural.

In order to do this, you should first decide whether the head of the noun phrase in question is a count noun (i.e., a noun which can take singular and plural forms, such as *cat*) or a mass noun (i.e., a noun which doesn't have a plural, such as *deer* or*furniture*). If the noun is count, use the values "plur" or "sing", depending on the noun in question. If the noun is mass, use the value "mass". With proper names, use "sing" for proper names denoting singular individuals, such as *John*, and "plur" with proper names indicating plural entities, such as *the Joneses*. Use "undersp-num" with noun phrases which are syntactically ambiguous between singular or plural, like collective noun phrases such as *the committee*. Use "unsure-num" if you are not sure whether the noun phrase is used as a count noun or a mass noun.

Tricky cases

- Conjoined NPs, plural quantifiers, and measure-NPs (*4 mg of Oestradiol*) should be marked with "plur" since plural pronouns are used to refer back to them.

- On the other hand, disjunctive NPs (*your doctor or your pharmacist*) are often "sing". This is particularly the case for when it's the head noun that is disjoint: *Your doctor or pharmacist will tell you what to do when she visits*.
- Nominalizations should be marked with "sing" because singular anaphoric pronouns are used to refer to them:

  Using Nerisone is not always easy. Doing this sometimes causes side effects

§18 The attribute ***Person*** specifies the (syntactic) person of a markable. Possible values are "per1" (for pronouns *I* and *we*), "per2" (for pronoun *you*), and "per3" (for any other noun phrase).

Notes

- This attribute is relatively easy to mark; the only problems arise in the case of coordination–e.g., for *you and me*, or *me and him*. In these cases, you should see which pronoun could be used to refer back to the NP, and use that pronoun's ***Person*** value. E.g., *we* is used to refer back to *you and me*, which should therefore be marked as "per1".

Grammatical function

Note: the values for this attribute used in ARRAU2 mainly come from the FrameNet annotation scheme. Some of the differences are that in FrameNet, "comp" and "np-compl" are conflated into one class; "comp" is defined in a slightly different way; and "predicate" is missing. The GNOME instructions for marking grammatical function were used in* arrau2*; they are repeated here. We are currently revising the scheme to make it compatible with the Universal Dependencies definition

This attribute is used to mark the grammatical function of an NP in the clause in which it occurs. For this reason, it should only be specified for NPs which occur in units with a verb. NPs that occur in parenthetical units which in turn occur inside NPs may be marked as "predicate" or "np-mod"; look at the explanation for these values below. All other NPs occurring in non-verbed units–in parentheticals not included in NPs, in titles, in listitems, etc.–should be classified as "no-gf". The attribute is marked by choosing the appropriate value from the menu next to the ***Gram_fnc*** attribute specification.

Attribute values

§19 "subj"

This value should be used for NPs that occur in subject position in verbed units, as in the following example:

```
And [Morris]subj followed very quickly after.
```

NPs should only be marked with this value if they have a subject.

§20 “obj” - the direct object

This value should be used for NPs in direct object position of transitive (such as buy) and ditransitive verbs (such as give):

> A new group of customers stimulated [the jewelry trade]$_{obj}$.
>
> The posthumous inventory of the French king Louis XIV’s possessions in 1720 describes [the table]$_{obj}$ in considerable detail.

This value should also be used with NPs occurring as non-subject arguments of a transitive phrasal verb (Quirk and Greenbaum, chapter 12). Phrasal verbs are verbs which consist of a verb plus a particle; they can be intransitive (as in sitting down) or transitive (like set up in John set up a new unit). The post-verbal argument of transitive phrasal verbs should be marked with “obj” rather than “adjunct” (the following examples are from Quirk and Greenbaum):

> We will set up [a new unit]$_{obj}$.
>
> Drink up [your milk]$_{obj}$ quickly.
>
> put on [a patch]$_{obj}$
>
> cover up [your patch]$_{obj}$

One test to recognize phrasal verbs is that in most cases the particle can either precede or follow the argument (except when the argument is a pronoun):

> They turned on [the light]$_{obj}$ / They turned [the light]$_{obj}$ on .
>
> They called off [the strike]$_{obj}$ / They called [the strike]$_{obj}$ off.
>
> He looked [it]$_{obj}$ up / * He looked up [it] .

for the purposes of this annotation, we are only going to consider as phrasal the verbs that pass this test; in all other cases, mark the NP as “adjunct”. On the other hand, when the main verb is *to be*, the values “predicate” or “there-obj” should be used; see next.

§21 “predicate”

Predicate is the grammatical function assigned to the complement in copula constructions, i.e., costructions in which the verb be is the head (as opposed to acting as an auxiliary), except in the case when the subject is expletive *there*, in which case the value “there-obj” should be used (see below).

> This is [a production watch]$_{predicate}$.
>
> the Palais-Royal was [the residence of the king’s cousin]$_{predicate}$

This value should also be used for NPs occurring in units marked as "paren-app":

`Anne-Marie Shillitoe,` [`an Edinburgh jeweller`]$_{predicate}$
`... inflammation around the mouth` [`perioral dermatitis`]$_{.}$

§22 "there-obj"

This value should be used for the post-copular NPs in ***there*** constructions, as in the following examples:

`There is` [`a man`]$_{there-obj}$ `in the garden`
`There is` [`considerable evidence`]$_{there-obj}$ `for the valuable, frequently gem-studded belts ...`

§23 "comp"

This is the value that should be used for NPs and PPs governed by a ditransitive verb such as ***give***, ***grant*** or ***allow***, which are not direct objects.

`The duke gave` [`the teapot`]$_{comp}$ `to my aunt`
`The king granted` [`him`]$_{comp}$ `the royal privilege of lodging in the Palais du Louvre`

§24 "adjunct"

This value should be used for all other NPs which occur as part of prepositional phrases inside of a unit (clause); if they occur inside of a NP, use "np-mod" or "np-compl" instead. These prepositional phrases may express the spatial or temporal location of the eventuality described by the verb, or instruments used to make some object, or the material:

`In` [`the courts of Europe`]$_{adjunct}$`, lavish quantities for formal diamond jewelry continued to be worn.`
`This jewel is made of` [`wood`]$_{adjunct}$`.`

Notice that PPs are not tagged in the scheme! Although in fact the PP is the adjunct, it is the embedded NP that is tagged as "adjunct", not the PP.

§25 "gen"

This value should be used for possessive NPs functioning as determiner:

[`its`]$_{gen}$ `mount`
[`the artist's`]$_{gen}$ `collection`

NPs occurring after an of particle should be classified as np-compl or np-mod even when they express possession::

the ring of [Jean de Grailly]$_{gen}$

§26 “np-compl”

This is the value to use for NPs occurring as post-nominal complements of an NP. Some of these complements can be recognized because they are indicated by the particle of:

the use of [acrylics]$_{np-compl}$
the design of [this chandelier]$_{np-compl}$
the straps of [a dress]$_{np-compl}$
the ring of [Jean de Grailly]$_{np-compl}$
Purple, white and green were the colours of [the suffragette movement]$_{np-compl}$

Notice that these NPs do not all have the same semantic function: in the second example above *design* clearly calls for an argument, whereas in the third, fourth and fifth example the particle *of* is used to express possession; however, we are going to mark all of these cases as “np-compl”. One distinction that we are going to make, however, is between these cases and partitive and quasi-partitive constructions, in which *of* is used to indicate the argument of a determiner rather than a noun. In these latter cases, “np-part” should be used instead (see next).

The more difficult cases are those in which the complement is not indicated by the particle *of*. Examples include:

the answer to [your question]$_{np-compl}$
the solution to [these problems]$_{np-compl}$

§27 “np-part”

This value should be assigned to NPs that specify the domain of quantification of a quantifier. These NPs also occur as arguments of the particle *of*, but the *of*-construction is used to specify an argument of the determiner rather than a noun:

Two of [them]$_{np-part}$ are buttons
Titanium is one of [the refractory metals]$_{np-part}$

In some cases, particularly noun phrases that refer to a certain amount of a given substance, the determiner / quantifier involves noun-like elements (as in *a lot*), making it difficult to decide whether “np-part” or “np-compl” should be used:

A lot of [effort]$_{np-part}$ went into the making of these early plastics
Three pounds of [bread]$_{np-part}$

Although the decision may be rather difficult in general, in some of these cases it is possible to decide by looking at whether the head noun of the embedding noun phrase indicates a measure; in these cases, “np-part” should be used.

§28 “np-mod”

This value should be used for NPs occuring in a PP which modifies a noun, but cannot be classified as either “np-compl” or “np-part”.

> the man with [a hat on [his head]$_{np-mod}$]$_{np-mod}$

This value should also be used for NPs included in parentheticals which occurs inside a NP:

> a clock of the same design and similar marquetry now in [the Ecole Nationale Superieure des Beaux-Arts]$_{np-mod}$

§29 “adj-mod”

For NPs that occur as arguments of adjectives:

> Are you allergic to [ny component]$_{adj-mod}$?
> the hanging oak branches are also typical of [Carlin’s work]$_{adj-mod}$
> if you are heavy with [water]$_{adj-mod}$

§30 “no-gf”

This value should be used for NPs which occur in units which are non-verbed and not a parenthetical or paren-app included in a NP (for which “np-mod” or “np-compl” should be used). These units include, among others, titles, list items, and parentheticals inside other units.

Possible difficulties

NPs which are part of a coordinated NP inherit their GF value from the grammatical function of the overall coordination.

Reference

Your next task is to annotate your interpretation of the markable. This begins by choosing a value for the attribute ***Reference***, that specifies the ***semantic type*** of the markable.

Types of noun phrases

Many, if not most, types of natural language expressions can refer anaphorically, but in ARRAU we are only interested in the anaphoric properties of ***noun phrases*** (NPs): expressions whose main word is a noun, like *the orange warehouse* or *a boxcar*, as well as proper names like *Bath* and pronouns like *it*. You will need to find which among these noun phrases are indeed anaphoric in the sense discussed above. One of your first tasks when analyzing a markable will be to identify what type of NP it is, using the attribute ***reference***.

Anaphoric noun phrases are an instance of ***term-denoting noun phrases***, so the first step towards identifying anaphoric expressions is to learn how to recognize these. A term-denoting NP is a noun phrase which is used to mention an object. The most typical examples of term-denoting NPs are proper names such as *Bath* or *Avon*, but many other types of noun phrases are term-denoting – including for example *two boxcars* in

> `There are [two boxcars] in Bath`

It is important however to realize that not all NPs are term-denoting. Your first interpretive task is to identify the semantic type of the noun phrase you are marking using the ***reference*** attribute. We first explain when each value should be used, then how they can be selected.

Idioms

Perhaps the clearest case of non-denoting NPs are NPs occuring in ***idioms*** such as *what [the heck]*. In this idiom, the NP *the heck* does not refer to anything – i.e., the meaning of the idiom is not derived from the meaning of the expression *the heck*. Other examples of idioms containing non-denoting NPs are *what [a pain in [the butt]]* (in which neither *a pain* nor *the butt* really refer to anything), *Kill [two birds] with [a stone]*, in which neither *two birds* nor *a stone* refer to anything, *Cut [John] [some slack]*, in which *John* is referring but *some slack* isn't, or *kicked [the bucket]* (in which *the bucket* does not refer to any object in particular. The value "idiom" of the ***reference*** attribute should be used when you think that a NP is part of an idiom.

Expletives

A second case of NPs that clearly do not denote terms are certain uses of the pronominal expressions *there* and *it*. These words ***can*** be (and often are) used to refer to objects, as in the following two examples:

> `My brother finally bought a dog. [It] is a big grey Alsatian.` (*it* = "the dog that my brother bought")
>
> `To meet Prof. Rodgers, go to the NLP Lab. [He] is often [there] at this time of the day.`
> (*he* = "Prof. Rodgers", *there* = "the NLP Lab")

However, in other cases, *it* and *there* only serve as 'placeholders' (in these cases, these words are called ***expletives***). In the first example below, *it* does not refer to anything: it is only there because of the syntactic requirement that English finite clauses need to have a subject even when the underlying predicate does not have an argument to be filled by this subject. *there* in the second example is there for the same reason.

> `[It] takes an hour to get to Corning`
>
> `[There] are two boxcars in Bath`

The value "expletive" of the ***reference*** attribute should be used for such cases.

Incomplete

A third category of non-referring expressions are the partial NPs that are the result of hesitations in spoken language, as in the following example from the Switchboard corpus:

> so, uh, with the issue of trial by jury, uh, I actually find the whole question about whether you need [a], [a unan-], a unanimous verdict in a criminal case to be somewhat interesting (Switchboard, 0043_4148)

These fragments of NP are treated in ARRAU as markables, but marked using the "incomplete" value of the ***reference*** attribute.

Predicates

In the cases above, the NP does not have a meaning at all. However, NPs may be non term-denoting even in some cases in which they do have a semantic meaning. For instance, in the sentence *it is a big grey Alsatian* seen above, the NP *a big grey Alsatian* does have a meaning, but it is not used to introduce a second entity in the discourse, and it does not function as an argument of a predicate. Rather, the function of the NP is to express a ***property*** of the dog that my brother bought. We call such NPs ***predicative***. Another example of predicative NP is *the best chess-player in the school* in *Hillary is* [*the best chess-player in the school*]. Again, the function of this NP is to ascribe a property to Hillary, not to introduce a new object in the discourse. In general, in many (although not all) ***copular*** clauses (clauses whose main verb is *to be*, such as the two examples just discussed) either the subject or the object is a noun phrase used to express a property.

Another costruction in which NPs often serve as predicates are ***appositions*** – non-restrictive nominal modifications typically expressed using parentheticals. An example of apposition is the NP *my nephew* in

> Carlo, [my nephew], is [a nice boy]

In this example, both NPs *my nephew* and *a nice boy* express predicates, and therefore are ***not*** term denoting, unlike *Carlo*, which introduces or refers anaphorically to an entity.

A third example of NPs typically interpreted as predications are NPs occurring as objects of verbs that specify that the entity in subject position is assuming a position, such as *become*:

> John became [CTO of the company]

The value "predicate" of the ***reference*** attribute should be used when you think that a NP is used predicatively.

As already said, the decision whether an NP is not predicative is not always easy, and cannot always be made on syntactic grounds alone. For instance, in

> [The Italian prime minister, [Antonio Conte]], arrived in London for meetings today.

it is the NP in appositive position (*Antonio Conte*) that acts as term-denoting, whereas the embedding NP has a predicative function. In so-called ***specificational*** copular clauses, it is the subject that is predicative, whereas the object is generally taken to be referential:

```
[The director of Anatomy of a Murder] is Otto Preminger
```

Whereas in so-called ***identificational*** copular clauses, both the subject and object are generally taken to be referring:

```
[That (woman)] is [Sylvia]
```

One test that you can use to decide whether an NP in one of the positions just discussed is predicative is to ask yourself whether the clause is talking about two objects or just one. This test works reasonably well, except in the case of identificational clauses, whose function is to equate two terms previously considered distinct. We consider both NPs in such clauses as referring, as well as the two NPs in so-called ***equative*** copular clauses which report the discovery that two entities hereto thought to be distinct are in fact the same:

```
We discovered that [the killer] was [Mr. Ray]
It has been claimed that [the location of the Battle of
Brunanburh] is in fact [Bromborough on the Wirral].
```

If you are marking a case of predication (an apposition or a copular clauses who you are sure is neither identificational nor equative), but you are not sure which NP is term-denoting and which one is predicative try to think which of the two could be viewed as a property of the other. When one of the NPs is a proper noun that is often referring, but this is not always the case. The proper name *Elisabeth II* is referring in the first and second of the following examples, but predicative in the third:

```
[Elizabeth II] is [the Queen of England]. [She] ...
[The Queen of England] is [Elizabeth II]. [She] ...
[She] was crowned [Elizabeth II] in 1953. [She] ...
```

if you find you can't decide do not worry too much, just make sure you only include one in the coreference chain.

Apart from the examples above, the NPs that are best viewed as expressing predicates or predicate modifiers are those that occur in locative expressions such as "to the left", or "go North", very common in some types of dialogue. "the left" and "North" are NPs, but in these examples the whole locative expression is best viewed as an adverbial not referring to any particular object. So you should not be concerned with marking such noun phrases as deictic or anaphoric, the only exception being cases of anaphoric expressions explicitly referring to directions.

Quantifiers

Another type of noun phrases that do not (always) denote terms is ***quantified*** NPs. A first example of quantified NPs are so-called ***wh-NPs***, such as *which engine* in

`and then [which engine] do you wanna use?`

What this example makes clear that the speaker is not referring to any object using the *wh-NPs*–on the contrary, he/she is trying to find out the engine that the other conversational participant(s) have in mind. The function of the *wh-NP* can be perhaps best understood by analogy with programming languages: *wh-NPs* can be viewed introducing a ***variable*** in the logical form of sentences, which has to be instantiated by an object in order to obtain a statement that can be evaluated (e.g., *I want to use engine E1*). By contrast, the stereotypical case of term-denoting NPs, proper nouns, can be viewed as specifying a ***constant***. Other examples of *wh-NPs* are *where* as in *where is the engine now* but also *how long* and *how many oranges*. By default, *wh-NPs* should be assigned a "quantifier" value for the ***reference*** attribute, except in the cases discussed below.

A second class of noun phrases that we treat as quantifiers are noun phrases such as *all of the boxcars* in

`[all of the boxcars] are empty`

In semantics, a distinction is made between two types of determiners. Determiners like *all*, *every* (as in *every boxcar*), *each*, *most*, *few*, and *no* do not have as a primary function to introduce or refer back to entities in a discourse (although they could do so indirectly, see below). Instead, they are viewed as expressing a relation between sets: they indicate the proportion of individuals in the first set that has the property characteristic of the objects in the second set. Thus, *all of the boxcars are empty* states that the set of boxcars is a subset of the objects that are empty–equivalenty, that all objects in the set of boxcars have the property of being empty. Again, these NPs can be viewed as introducing variables in the logical form of the sentence, which can be interpreted as 'take any x such that x is a boxcar. then that x is going to be empty'. For another example, *no boxcars* in the following example states that the intersection of the set of boxcars and the set of empty objects is the null set–this can be rephrased in terms of variables as meaning: 'take any x such that x is a boxcar. then that x is NOT going to be empty'.

`[no boxcars] are empty`

Again, by default NPs that introduce such quantifier should be assigned a "quantifier" value for the ***reference*** attribute, except in the cases discussed below.

Other determiners do not behave this way, however. Indefinite NPs (NPs with the determiners *a*, *a few* and *some* (like *an engine*, *someone* or *a few towns*), NPs with numerical determiners, such as *two boxcars* and *five engines*), and definite NPs such as *the engine* or *both boxcars*, are generally considered term-denoting. So, for example, in *send [two engines] to*

Avon, *two engines* is not generally considered a quantifier. There is a lot of debate in the literature regarding which NPs should be considered quantifiers and which NPs should be considered as term-denoting; in ARRAU

- We consider indefinite NPs (including NPs with phrasal determiners beginning with *a*, such as *a lot of*); definite NPs; numerical NPs; and NPs with the determiners *many*, as term denoting. (For those who are familiar with typed-logical frameworks, these are all NPs considered to be of type e.)
- We consider *wh-NPs*; and NPs with determiners *all*, *every* and *each*; *most*; *few*; *no*; and *any* (as in *any of the boxcars* but also *anything*); as quantifiers.

Anaphoric reference to quantifiers is possible, but it follows different rules from anaphoric reference to terms–in particular, 'direct' reference is generally not possible outside the scope of the quantifier:

> [Every student] entered the classroom. *[He/She] was happy.

Two types of reference are however possible. First of all, it is possible to refer to the variable that, as explained above, is implicitly introduced in the logical form of a sentence by a quantifier. Such ***bound anaphoric references*** have values that vary with the instantiations of the variable, again just like variables in a function.

> [Every student]$_i$ sat at [his/her]$_i$ desk.

Assume three people are relevant here: Bill, Mary and Sue. The meaning of this sentence is that if you take any x in that set, that x sat at x's desk: Bill sat at Bill's desk, Mary sat at Mary's desk, and Sue sat at Sue's desk.

A second case of anaphoric reference to quantifiers are plural references to the set of entities for which the predication in the sentence containing the quantifier holds. E.g., in the following example, *they* refers to the majority of students that were happy.

> [Most students] entered the room. [They] were happy.

Both in the cases of bound anaphora and of plural reference to quantifiers, we follow a different approach. Instead of marking the quantified NP as a quantifier, thus non-referring, we mark it as "new", and use the ***generic*** quantifier to specify that it is a quantifier, as we will explain below.

Finally, a quantified NP may sometimes be indirectly anaphoric in that the set quantified over (the ***domain of quantification***) has already been introduced. Consider the following variants of the example above, where a set of students is introduced in the first sentence:

> [The students] waited in the corridor until the teacher arrived.
> Then [all of [the students] entered the classroom.]
> Then [every student] entered the classroom.

In these examples, the domain of quantification is the set denoted by *the students* in the first sentence. In the first sentence, the domain of quantification for *all of the students* is indicated explicitly by a second NP. In the second case there is no explicit markable for the domain of quantification. One should therefore be introduced:

> [The students] waited in the corridor until the teacher arrived.
> Then [every [student]] entered the classroom.

Coordination

Coordinated NP can provide antecedents for singular and plural anaphoric reference and are therefore treated as markables in ARRAU.

> [[John] and [Mary]] will arrive soon. [They] are never late.
> [[John] and [Mary]] will arrive soon. [She] is never late.
> [[John] or [Mary]] will arrive soon. [They] are never late.

In ARRAU however we do not treat coordinated NP as term denoting, as we use the ***multiple antecedents*** mechanism discussed below to annotate in the same way both the cases of plural reference just discussed and cases in which the antecedents are separated:

> [John] will be here soon with [Mary]. [They] are never late.

Coordinated NPs should thus be treated as non-referring; the "coordination" value should be used for the ***reference*** attribute. (Coordinated NPs are also special in that they are the one type of NP for which we do not mark a head–***min_ words***–as explained earlier.)

Undef-reference

The "undef-reference" value should be used for noun phrases which are term-denoting, but whose interpretation is vague or not specified. Examples include so-called ***generic (or situational) they*** cases–cases of anaphoric reference to some unspecified entity in the situation, as in the following example, where pronoun *they* refers to one or more individuals in the restaurant that are not specified.

> Sue entered the restaurant, but [they] told her that Kim had left.

Another examples marked as undef-reference are uses of *one* or *you* to indicate some unspecified individual:

> so, uh, with the issue of trial by jury, uh, I actually find the whole question about whether [you] need a, a unan-, a unanimous verdict in a criminal case to be somewhat interesting
> (Switchboard 0043_4148)

How to annotate the semantic type of markables

§31 The attribute ***Reference*** is used to specify the semantic type of markables according to the guidelines just discussed. This attribute is marked at two different levels.

At the higher level, one of the following values can be specified: "new", "old", "non_referring", and "undef_reference".

We will explain in this section how to choose the appropriate value. As you'll see, when you choose one of the values "new", "old", or "non_referring" additional attributes appear in the window. We will explain now what to do in case you have chosen "non_referring," and what to do in the other two cases in the next sections.

§32 The values for the ***Reference*** attribute have the following interpretation:

old: A term-denoting markable in the sense discussed above, i.e., which refers to a concrete or abstract object which has already been mentioned or discussed earlier in the dialogue.

For example, in the following excerpt, the markable *it* refers to an object already mentioned using the phrase *a boxcar*.

```
M: [I] want [you] to take [a boxcar] from [Elmira]
and load [it] with [oranges]
```

Examples of abstract objects are facts, events, actions, and plans. For instance, in the example just shown, M is proposing a plan: to take a boxcar from Elmira and loading it with oranges.

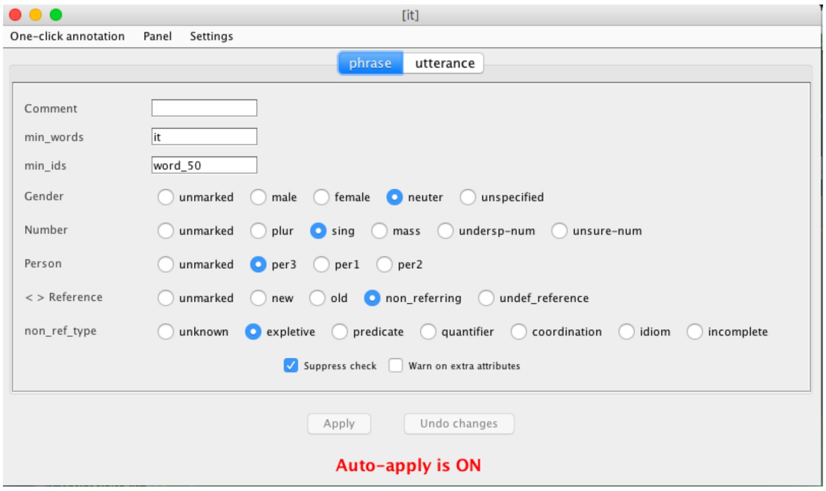
Note that the previous mention of the object need not have been made with a markable! (See discussion of ***segment*** below.)

We will explain below how you can indicate which previously mentioned object is being referred to in this case.

new: A term-denoting markable which in your opinion is the first mention in the dialogue of a concrete or abstract object. This value should also be used for markables that you want to mark as referring to an object associated with a previously mentioned object (see below).

undef_reference: A term-denoting markable whose interpretation is however unspecified in the sense discussed above–e.g., a generic *you*.

non_referring: A markable which does not refer to an object, whether concrete or abstract. If you choose this value, a new attribute, ***non_ref_type***, will appear, as illustrated by the following screeenshot



Comment

min_words it

min_ids word_50

Gender ○ unmarked ○ male ○ female ● neuter ○ unspecified

Number ○ unmarked ○ plur ● sing ○ mass ○ undersp-num ○ unsure-num

Person ○ unmarked ● per3 ○ per1 ○ per2

< > Reference ○ unmarked ○ new ○ old ● non_referring ○ undef_reference

non_ref_type ○ unknown ● expletive ○ predicate ○ quantifier ○ coordination ○ idiom ○ incomplete

☑ Suppress check □ Warn on extra attributes

Auto-apply is ON

and you will be requested to indicate why you think this markable is non-referring by choosing among the classes of non-referring markables discussed above.

(a) The markable is the word *there* or *it* used as placeholder.

`[There] are two boxcars in Bath`

(= "two boxcars are in Bath")

`[It] takes an hour to get to Corning`

(= "to get to Corning takes an hour")

Choose the value "expletive" if you think this is the interpretation. (Keep in mind however that other uses of *there* and *it* do refer to concrete or abstract objects, and should be marked as "old" or "new" as appropriate!)

(b) The markable is a noun phrase used as a predicate, as discussed above:

`The boxcar in Dansville is [a relic]`

Use the value "predicate" in this case.

(c) The markable is a noun phrase used as a quantifier, as discussed above.

`Are there [any boxcars] in Dansville?`

Use the value "quantifier" in this case, unless you see a bound anaphoric reference or a plural reference to the quantifier in the following text.

(d) The markable is part of an idiom, as in *what* [*the heck*]. Use the value "idiom" in this case.

(e) The markable is an incomplete fragment, as in the [*a*], [*a unan-*] example see above. Use the value "incomplete" in this case.

Do not leave any markables with the value "unmarked" – its purpose is to serve as an indication that an appropriate value has not yet been selected.

Marking anaphoric and deictic reference

If you decide that a markable is referring ("new" or "old"), you'll be asked to provide more information about its reference. In this section you'll find instructions to do this.

General principles

As said in the Introduction, our primary interest is the study of linguistic and deictic entity context-dependence. Thus, one of your most important tasks will be to identify cases of anaphora in the general sense discussed above: i.e., expressions referring directly or indirectly to an object already mentioned. We already saw a few examples of expressions of this type, like the pronoun *it* in the following example:

```
M: [I] want [you] to take [a boxcar] from [Elmira] and load
[it] with [oranges]
```

These cases should be marked by first choosing "old" as the value of the ***Reference*** attribute using the criteria discussed earlier, then proceeding to mark the markable as a case of "phrase" reference as discussed below.

The most important principle to keep in mind when annotating "old", "phrase" markables is that we aim to mark *all* cases of identity coreference, not just cases of anaphoric reference with pronouns and not just cases of coreference between specific mentions of entities; e.g., we also want to mark cases of generic coreference, of bound anaphora, etc.

In this vein of marking all cases of identify reference, we also consider as anaphoric those expressions that refer to an abstract object such as a plan or an action. Such abstract objects may have been previously mentioned using a phrase (a markable), or using a ***segment*** – one or more utterances describing the plan, as in the following example, where *this* refers to the plan in the previous utterance by M:

```
M: Take [engine E1] to get [the boxcar] to Elmira.
S: All right, [this] will take one hour
```

In addition, you will also be asked to mark cases of ***direct reference*** to entities in the surrounding situation or the wider world. The simplest example of direct reference are deictic references to to landmarks in the map in the TRAINS domain. For example, in the following utterance, both *engine E2* and *Corning* are objects on the map shown in Figure 1.

```
M: send [engine E2] off with [a boxcar] to [Corning] to
pick up [oranges]
```

In addition to subsequent mentions of objects already introduced in the discourse and deictic references to objects in the map you will also be requested to mark bridging references in the sense discussed above – references to objects which have not yet been mentioned, but are strictly related to objects which already have, as in the case of *the wheel* referring to a part of *the boxcar at Elmira* above. These cases are relatively rare in the TRAINS dialogues, but they are much more common in other genres in the ARRAU corpus. In fact, they are so common that we will not be able to annotate all such cases, and a big part of our instructions is concerned with restricting the range of what we annotate as bridiging.

Marking instructions: "new"

§33 Just as in the case of "non_referring", when you select "new" as the value for the ***Reference*** attribute, other attributes will appear.

§34 The attribute (Semantic) ***Category*** should be used to mark the type of object the markable refers to, if any. The possible values for this attribute and instructions for its annotation are provided in the next subsection. The ***Category*** attribute will also appear when you choose "old" as the value for ***Reference***.

§35 The next attribute is used to specify whether the markable is directly referring or not. This is done differently in the TRAINS domain and in the other domains. In TRAINS, where most of the objects that can be referred to are visible on the map in Figure 1, you will see an attribute called ***On_ map***. If you set this attribute to the value "yes", a second attribute will appear, called ***Object***, which allows you to specify (using a menu) which object the markable refers to. In all other domains, you will only see an ***Object*** attribute, but without a menu: you will have to write the name of the object the markable refers to. Instructions on how to mark direct reference are provided in this Section after the instructions for "old" markables.

§36 The next attribute, ***related_ object***, is to be set to "yes" when the markable is a bridging reference in the sense discussed above. The instructions for annotating this attribute are found later in this Section, after the instructions for marking direct reference.

§37 The last attribute you will see appear when you annotate a markable as referring is ***Genericity*** - the instructions for this attribute are after the instructions for ***Category***.

Marking instructions: "old"

§38 If you select "old" as the value for ***Reference*** further attributes will also appear, each on a separate line. Some of these attributes are the same that you see appear when you choose "new", but other attributes are used to specify which previously mentioned object the markable refers to.

§39 The ***Category***, ***Object*** (or ***On_ map***), and ***Genericity*** attributes are the same as for markables annotated as "new".

§40 The ***Ref_ type*** attribute is used to specify the ***reference type*** of the markable. Possible values are:

phrase: Use this value if the markable refers to an object which was already mentioned using a markable. E.g.,

> M: I want you to take [a boxcar] from [Elmira] and
> load [it] with [oranges]

it refers to the same object as *a boxcar*

segment: Use this value if the markable is ***discourse deictic***–i.e., if the markable refers to an abstract object – for instance a plan, event, action, or fact – which was discussed in an earlier segment of the dialogue, but not referred to using a markable.

> M: Take [engine E1] to get [the boxcar] to Elmira.
> S: All right, [this] will take one hour

this refers to the plan in the previous utterance by M

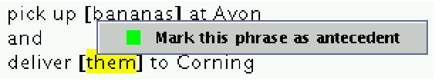
Note: subsequent references to the same abstract object should be marked as "phrase", since the object discussed in the segment has now been mentioned using a markable.

§41 If you select "phrase," then you should also indicate the most recent mention of the object that the markable refers to.

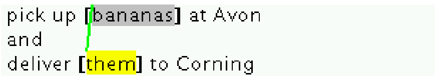
(a) In the dialogue window, right-click (using the right mouse button) on the brackets that surround the phrase which the current markable refers to.

(b) The text "Mark this phrase as antecedent" will appear; click on it with the left mouse button to set the mark.

This will cause a number to appear in the attribute window next to the "Single_phrase_antecedent" attribute, and will also cause a green line to be drawn in the dialogue window between the markable and the phrase it refers to.



pick up [bananas] at Avon
and
deliver [them] to Corning



pick up [bananas] at Avon
and
deliver [them] to Corning

As discussed above, you should mark *all* cases of entity coreference as "old", "phrase". These include, in addition to the examples of pronominal reference to specific mentions seen above:

- cases when the anaphor is not a pronoun, as in the last mention of Atco in the following example from ARRAU:

  > [*Atco Ltd.*] said [its] utilities arm is considering building new electric power plants, some valued at more than one billion Canadian dollars (US$851 million), in Great Britain and elsewhere. .... C.S. Richardson, [Atco's] senior vice president, finance, said ....

- cases of reference when the entity referred to is generic:

  > [*Solo woodwind players*] have to be creative if [they] want to work a lot, because [their] repertoire and audience appeal are limited.

- cases of bound anaphora where the antecedent is a quantifier:

Montedison S.p.A. definitively agreed to buy [*all of the publicly held shares of Erbamont N.V.*] for [$37 each].

- Cases of entity coreference between entities mentioned as prenominal modifiers, such as *stock market* in the following example:

  .. [small investors] seem to be adapting to [greater [*stock market*] volatility] . . . [Glenn Britta] . . . is ''factoring'' [[the market's] volatility] ''into [investment decisions].'' (*Amir Zeldes example from the ONTONOTES corpus*)

This also includes cases when the anaphor is a plural reference to *more than one antecedent*, as discussed in the Section on 'aplit antecedent' plurals below.

§42 If a markable refers to an object that was mentioned earlier in the dialogue more than once, mark a reference to the most recent mention.

the second *it* points back to the first *it*

M: I want you to take [a boxcar] from [Elmira] and load [it] with [oranges], then send [it] to [Corning]

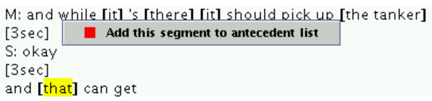
§43 If you select "segment", then you should also indicate which text region the markable refers to. You do this by selecting the appropriate unit(s) of text.

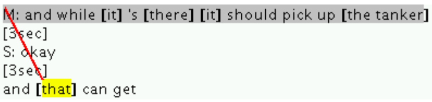
(a) In the dialogue window, right-click (using the right mouse button) a word contained in a unit which is part of the text that evokes the abstract object that the markable refers to.

(b) MMAX will show you the units that overlap with the position that you clicked. Choose the appropriate unit. text "Add this segment to antecedent list" will appear; click on it with the left mouse button to set the mark.

(c) Repeat this process until you have added all the desired units.

This will cause a series of numbers to appear in the attribute window next to the "Segment_antecedent" attribute, and will also cause a red line to be drawn in the dialogue window between the markable and each of the selected lines.


M: and while [it] 's [there] [it] should pick up [the tanker]
[3sec]
S: okay
[3sec]
and [that] can get


M: and while [it] 's [there] [it] should pick up [the tanker]
[3sec]
S: okay
[3sec]
and [that] can get

Examples of anaphoric references that should be marked as "segment" include:

- In TRAINS, cases of reference to (parts of) a plan, as seen in the example before:

the second *it* points back to the first *it*

M: Take [engine E1] to get [the boxcar] to Elmira.
S: All right, [this] will take one hour

- In any domain, cases of ***event anaphora***–i.e., reference to an event that actually occurred–as in the following artificial example:

  > John arrived late. [That] upset Mary.

- There are also however more complex cases of discourse deixis, as in the following example, where *the problem* does not refer to an event, but to a problem originated by the limitations of woodwind players' repertoire:

  > Solo woodwind players have to be creative if they want to work a lot, because their repertoire and audience appeal are limited.
  > The oboist Heinz Holliger has taken a hard line about [the problem] ....

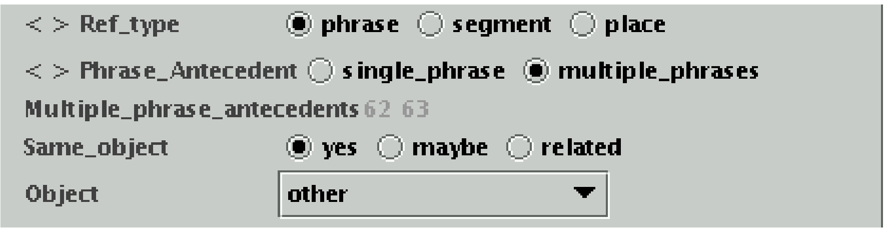
Split antecedent plurals

By default you can only mark one reference type, and if you choose "old" then you can only set a pointer to a single phrase or segment. Sometimes, however, you'll find that a markable may refer to more than one object previously only mentioned by distinct markables. In this Section we provide instructions for this situation. We discuss below a different situation: the case in which the markable is ***ambiguous***, i.e., it has more than one interpretation.

§44 A ***plural*** markable is one which refers to a set of objects already mentioned using a phrase. For example, the markable *them* in the following snippet refers to the combination of the engine and the boxcar.

> S: Please hook up [the engine] and [the boxcar], and send [them] to Elmira.

In these cases, when all elements of a set are introduced, a special mechanism to mark more than one pointer should be used instead of marking this as a case of bridging reference (discussed below). In the example above, two pointers from *them* should be marked – one to *the engine* and one to *the boxcar*. To set multiple ***phrase*** pointers, select the "multiple_phrases" button.



< > Ref_type ● phrase ○ segment ○ place
< > Phrase_Antecedent ○ single_phrase ● multiple_phrases
Multiple_phrase_antecedents 62 63
Same_object ● yes ○ maybe ○ related
Object other ▼

Only mark multiple antecedents this way if they refer to distinct objects. If two possible antecedents refer to the same object, just mark the most recent one.

Marking instructions for direct reference (the *Object* and *On_map* attributes)

The ***Object*** and ***On_map*** attributes should be used to specify markables that directly refer to an entity in the world, whether an entity in the visual situation as in the cases of deixis, or in the more general 'larger situation'.

In domains where the text is interpreted without reference to an explicity visual situation, such as RST, we take the world to be our situation of interpretation, and will mark as directly referring all mentions of entities in the world using proper names, such as *Atco Ltd.* and *Great Britain* in the following snippet:

```
[Atco Ltd.] said its utilities arm is considering
building new electric power plants, some valued at more
than one billion Canadian dollars (US$851 million), in
[Great Britain] and elsewhere.
```

Other markables that we consider directly referring are: first and second person pronouns (*I*, *us*, *your*); mentions of dates (*1986*, *November 3rd 2021*); and definite descriptions referring to unique entities in the world and that therefore act essentially as proper names (*the pope*, *the film ''Klute''*).

§45 You begin to mark reference by choosing an appropriate value for the ***Reference***. If you annotate the markable as referring, i.e., if you choose the values "new" or "old" for this attribute, the attributes for marking direct reference appropriate for the domain appear, and you will then be able to specify the object referred to by the expression you are marking.

§46 Direct reference is marked differently in the TRAINS domain (which is a closed world with a small number of objects) and in the other domains. (Note: whether you need to mark direct reference or not after the first mention of an entity also depends on the domain, see below.)

§47 In TRAINS, you will see an attribute called ***On_map***. If you set this attribute to the value "yes", a second attribute will appear, called ***Object***, which allows you to specify (using a menu) which object the markable refers to. (You'll find that deictic references to towns using proper names, such as *Corning*, have already been automatically marked by us, but there are a few cases of references to towns – e.g., using pronoun *there* – that haven't been.) The menu contains all the objects on the map, but these are not all the objects discussed in the dialogue. If the markable clearly refers to one of the objects on the map, select that object from the menu; if the markable refers to an object which is not on the map, or it is unclear which of the objects on the map it refers to, select "other" (which is also the default value). You should also use "other" for all deictic references to the speaker using first and second person pronouns.
Note: in TRAINS, you mark the object referred to for each directly referring mention.

§48 In all other domains, where there is no predefined list of objects, you will directly see the ***Object*** attribute, but without a menu. So, the first time a markable mentions an object, you will have to write the name of the object the markable refers to. Ideally you should provide a fairly complete name for the object so that we may later try to link the mention e.g., to Wikipedia, but the important thing is that you provide a name so we know the mention is directly referring.
Note: you only need to do this the first time around; after that you simply link the markable to previous mentions of the same entity using the anaphoric reference mechanisms discussed above.
Note: this is also how we proceed for other dialogue domains such as LIGHT even though the list of objects in that domain is also predefined.

Marking instructions for the *Related_object* attribute (bridging references and other cases of coherence expressed via non-identity relations)

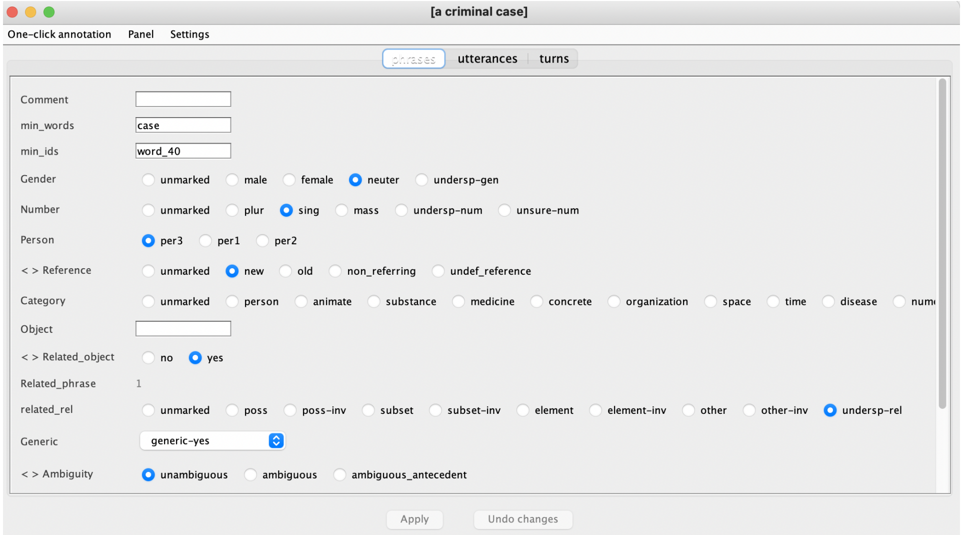
The attribute ***Related_object*** is used to indicate markables that establish entity coherence via non-identity relations. This includes, first of all, bridging references, also called ***associative references***–a type of anaphoric reference which links the object being referred to by the markable to an *already established* discourse entity (which in this case is called the ***anchor***) via a semantic relation other than coreference, as in the 'wheel' example seen earlier, repeated here for convenience:

```
S: Bad news about [the boxcar at Elmira]. [They] tell me
[the wheel] is broken and will have to be fixed.
```

This type of anaphoric reference is called 'bridging' because it involves a 'bridging inference' based on lexical or encyclopedic knowledge (in this case, knowing that wheels are parts of cars). The stereotypical associative bridging reference is an NP with a ***functional*** head noun–a noun which has an implicit argument. E.g., in the example above, *the wheel* really is a sort of abbreviation for *the wheel of the boxcar at Elmira*. The first question you should ask yourself is if the markable you are considering has such an implicit argument. If so, the markable is a strong candidate for a bridging reference; you should then test if the bridging reference satisfies the second constraint discussed below, namely, that the reference is required to establish a relation of entity coherence between two distinct clauses. Note however that, as we will see below, we also ask you to mark as 'related object' markables that, while not functional, establish entity coherence.

In fact, we also use the attribute ***Related_object*** to mark other types of indirect anaphoric reference not generally assumed to involve the establishment of a semantic relation between the anaphor and the anchor via a bridging inference, as in the following cases, where again the underlined NP is linked by a non-identity relation to its anchor, but the relation is supplied by the semantics of the anaphor:

```
[John] bought [a red t-shirt], and [Bill] bought
[a green one]. (identity of sense anaphora)
```



Comment
min_words case
min_ids word_40
Gender ○ unmarked ○ male ○ female ● neuter ○ undersp-gen
Number ○ unmarked ○ plur ● sing ○ mass ○ undersp-num ○ unsure-num
Person ● per3 ○ per1 ○ per2
< > Reference ○ unmarked ● new ○ old ○ non_referring ○ undef_reference
Category ○ unmarked ○ person ○ animate ○ substance ○ medicine ○ concrete ○ organization ○ space ○ time ○ disease ○ num
Object
< > Related_object ○ no ● yes
Related_phrase 1
related_rel ○ unmarked ○ poss ○ poss-inv ○ subset ○ subset-inv ○ element ○ element-inv ○ other ○ other-inv ● undersp-rel
Generic generic-yes
< > Ambiguity ● unambiguous ○ ambiguous ○ ambiguous_antecedent

Figure 7: The attributes ***Related_phrase*** and ***related_rel*** that appear in the MMAX window when ***Related_object*** is set to true

> [One reason] to drive [electric cars] is to do [[your] bit] in [the fight against [climate change]]. [Another reason] is that [they] are much cheaper to run. (*other-* anaphora)

In the annotation of this attribute we are trying to find a balance between two conflicting objectives. On the one end, we want to make sure that all cases of entity coherence in our texts are covered. On the other end, almost every markable in a coherent text will be connected to entities in the context in some way; we cannot mark everything, so we need to limit the scope of the annotation in some way. To this end, we have adopted the following restrictions.

First, we only mark relations that establish a connection between separate discourse units / clauses. Thus, although we mark possession, we do not mark a relation between the speaker and the car in the following example, where both entities are arguments of the same predication:

> [I] have [a car]

Second, we only mark a connection to one anchor. Thus, in the following example, we only mark *a bolt* as being a bridging reference to *the wheel*, not also to *the car*.

> [I] almost crashed [[my] car] yesterday.
> [The wheel] came off, because [a bolt] had loosened.

More detailed instructions follow.

§49 The attribute ***Related_object*** is to be set to "yes" when the markable establishes entity coherence with a *previous* utterance by referring to an entity related to an entity introduced in that utterance, in the sense discussed above. Setting the attribute to "yes" results in several other attributes to appear in the MMAX window, as shown in Figure 7.

§50 The first thing to do after setting ***Related_object*** to "yes" is to indicate the most recent mention of an object to which the reference of the markable is related, as follows.

(a) In the dialogue window, right-click (using the right mouse button) on the brackets that surround the phrase denoting the related object.

(b) The text "Mark this phrase as related" will appear; click on it with the left mouse button to set the mark.

This will cause a number to appear as value for the ***Related_phrase*** attribute, and will also cause an orange line to be drawn in the dialogue window between the markable and the phrase it refers to.

In case the anchor is mentioned multiple times, please try to mark the last mention as related.

§51 Setting ***Related_object*** to "yes" will also result in a second attribute appearing, called ***related_rel***. This is how the attribute should be used to specify the types of related/associative references we ask you to mark:

part and generalized possession relations The bridging reference to the wheel in the example seen earlier is an example of markable referring to an object that stands in a ***part-of*** relation to an object previously mentioned. You should mark these markables as ***Related_object*** = "yes" using the instructions below, and then specify the relation as "poss". You should also use the "poss" value for all cases of bridging reference expressing some form of possession relation, as in the following example, where *the first half* is a reference to a part of a concert:

> **[***Richard Stolzman's recent appearance at the Metropolitan Museum***]** ... was a case in point. In **[**<u>the first half</u>**]**, he ...

you also use "poss" to mark bridging references expressing ***attributes*** of entities, as in the following case, where *the income* refers to Kellogg's income for the year:

> **[***Kellogg***]** reported its financial results for the year yesterday. **[**<u>The income</u>**]** grew to ....

The "poss" value is used when, as in the examples above, the bridging reference refers to a part or attribute of an entity alredy introduced in the discourse. In some cases, however, the part is introduced first, and the possessor second. In these cases, you should use the value "poss-inv", as in the following example:

> **[***The handle***]** slowly turned, and **[**<u>the door</u>**]** started opening.

set relations The second broad class of bridging relations of interest are those that hold between a set and its elements, or between a set and a subset. For instance, in the following example:

> There are [*two boxcars*] at Bath.
> We should send [one] (or: [one boxcar]) to Avon

Both [*one*] and [*one boxcar*] are references to elements of the set introduced in the previous utterance. Again, these markables are examples of bridging references that you should mark as ***Related***, then choosing the value "element" for the ***related_rel*** attribute.

Bridging references can also express ***subset*** relations, as in the following example:

> We have [*three engines in tota*].
> [Two] are at Elmira.

in this case, the markable [*Two*] refers to a subset of the set of three engines mentioned in the first utterance, and again should be marked as related, specifying the value "subset" for the ***related_rel*** attribute.

"element" and "subset" should also be used to indicate relations between ***types*** and their ***instances***. In the following example, the markable *The oboist Heinz Holliger* refers to an instance of the type *Solo woodwind players* (which is also referred to using plural pronouns *they* and *their*–incidentally, remember that the latest mention should be marked as related phrase), and the value "element" should be used for the ***related_rel*** attribute.

> [*Solo woodwind players*] have to be creative if
> [*they*] want to work a lot, because [*their*]
> repertoire and audience appeal are limited.
> [The oboist Heinz Holliger] has taken a hard line
> about the problem ....

or between types and ***subtypes***, as in the following example, where In the following example, the markable *Shostakovich quartets* refers to a subtype of the type *chamber music*, and the value "subset" should be used for the ***related_rel*** attribute.

> Managers and presenters insist that [*chamber*
> *music*] concerts are a hard sell, but can audiences
> really enjoy them only if the music is purged of
> threatening elements, served up in bite-sized
> morsels and accompanied by visuals?
> What's next?
> Slides to illustrate [Shostakovich quartets]?

As with "poss" relations, inverse relations can also occur; in this case, the -inverse version of the attributes should be used.

other anaphora A third case of relatedness we want to mark are expressions containing the word *other* and referring to a second object of the same type as an object already mentioned. E.g., in the TRAINS dialogues it is common to first talk about one engine, then introduce a second engine, and then return talking about the first engine by saying *the other engine*. These cases, as well, should be marked as ***Related***, and the value "other" should be specified for ***related_rel***.

undersp-rel Finally, there are a number of clear cases of bridging reference and/or semantically expressed relatedness in which the markable clearly serves to establish entity coherence with an entity in a previous utterance through a non-identity relation, but that do not fall into any of the categories above. These include, for instance, the identity of sense cases discussed above:

> [John] bought [a red t-shirt], and [Bill] bought
> [a green one]. (*identity of sense* anaphora)

as well as any other case in which the markable clearly expresses a bridging reference–e.g., cases of so-called 'situational' reference as in the first example below, or cases in which the head noun is clearly functional, but the relation to the anchor does not clearly fall in the categories of generalized possession or set relations discussed above:

> [*The dance*] had started.
> [The orchestra] was playing.
> [*John Smith*], died February 3rd.
> [Parents] unknown.

You should mark these cases using the "undersp-rel" value.

§52 Notice that relatedness in the sense used here is not only expressed using definite NPs. First of all, indefinite NPs can be functional and express entity coherence as well, as in the following example:

> [*Kellogg*] reported its financial results for the year
> yesterday. [Income] grew to ....

In fact, we are also asking you to mark as ***Related_object***="yes" cases in which the markable is not relational, but the reference nevertheless establishes entity coherence with a previous unit, as in the following example, where both *Madrid* and *Barcelona* establish a relation of entity coherence with the entity Spain.

> I went to [*Spain*] last month.
> I first stopped in [Madrid], then I went to
> [Barcelona].

§53 In most cases, the bridging reference refers to an object not previously mentioned in the dialogue. In these cases, you should first choose "new" as the value of ***Reference***, then specify that the markable is "Related" to a previous object. After you've done this, you'll be able to identify the antecedent using a pointer.

In a few cases, you'll find that the bridging reference refers to an object which has already been mentioned. In these cases, you should always choose the value "old" for the markable, and identify the antecedent. However, you should also set the ***Related_object*** value to yes if the antecedent is not in the previous utterance, as in this case we would otherwise miss a case of (associative) entity coherence.

§54 There are also a number of cases you should make sure *not* to mark as ***Related***. These include:

- Cases in which the anchor is explicitly provided as part of the markable, as in the following example, where the fact that the income is Kellogg's is explicitly stated in the markable. In such cases, only mark the (identity) anaphor *the company* as "old".

  [*Kellogg*] reported its financial results for the year yesterday. [The [company's] income] grew to ....

- Cases in which the relation is explicitly provided by the clause containing the markable you are annotating and the candidate anchor. In the following example, it is the text that explicitly states that there is a relation of ownership between the velvet suit and Richard Stolzman; it doesn't have to be annotated.

  [Richard Stolzman] was clad in [a trademark velvet suit].

  Other semantic relations which are explicitly stated by the text include the relations expressed by possessive constructions, such as the relation between E1 and its boxcar in *E1's boxcar*. Do not mark such relations.

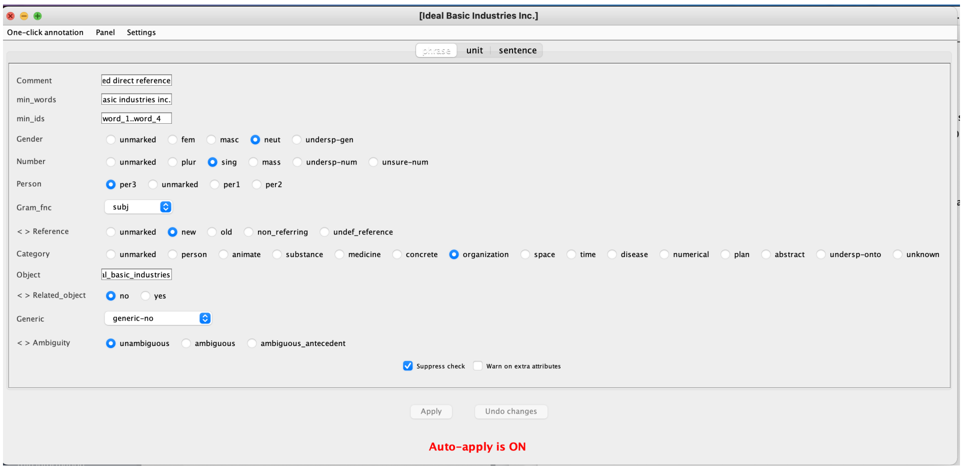
- Many bridging references are related to multiple anchors, but do not annotate more than one.
- Do not use the ***Related_object*** attribute to mark the cases in which a plural anaphoric expression is used to refer to a set of objects introduced singularly before, as in *Kim saw Robin. They had been good friends at school.* In this case, mark *they* as a plural reference using the methods discussed earlier.

Semantic category

After choosing the appropriate value for the ***Reference*** attribute, if you classified the mention as referring by choosing "new" or "old" as its value you will be asked to mark the type of object the mention refers to using the ***Category*** attribute, as illustrated in Figure 8.

General Principles

As we are primarily interested in anaphoric reference, one guiding criterion of this aspect of the annotation is (i) to annotate semantic distinctions that are known to affect anaphoric interpretation. In particular, we classify entities into three high level categories: animate entities, other concrete entities, and abstract objects. Another criterion is (ii) we use underspecified types to avoid making choices when not necessary.



Comment: ed direct reference

min_words: asic industries inc.

min_ids: word_1..word_4

Gender: ○ unmarked ○ fem ○ masc ● neut ○ undersp-gen

Number: ○ unmarked ○ plur ● sing ○ mass ○ undersp-num ○ unsure-num

Person: ● per3 ○ unmarked ○ per1 ○ per2

Gram_fnc: subj

< > Reference: ○ unmarked ● new ○ old ○ non_referring ○ undef_reference

Category: ○ unmarked ○ person ○ animate ○ substance ○ medicine ○ concrete ● organization ○ space ○ time ○ disease ○ numerical ○ plan ○ abstract ○ undersp-onto ○ unknown

Object: al_basic_industries

< > Related_object: ● no ○ yes

Generic: generic-no

< > Ambiguity: ● unambiguous ○ ambiguous ○ ambiguous_antecedent

☑ Suppress check □ Warn on extra attributes

Auto-apply is ON

Figure 8: The ***Category*** attribute in the MMAX window and its attributes

We make a distinction between two types of animate entities, people (for which we use the category "person") and all other animate entities (e.g., animals), for which the value "animate" is used. Concrete entities are physical objects that can be touched, such as people, animals, houses, and trees. We use specific categories for some of these entities: besides people and other animated entities as discussed above, we also treat specially substances (e.g., gold, water), and medicines. For all other concrete entities (e.g., works of art, buildings, trees) the more general "concrete" category should be used. A category of entities which could be considered either concrete or abstract are spatial locations; for these we would use a separate category as well. "space". The category "organization" should be used for references to institutions, whether used in an animate or in an inanimate sense. Every other entity is considered abstract in some sense. Among these entities we use special categories for temporal expressions (*this century*), numerical expressions used referentially, and expressions referring to actions and events (*the second world war*); all other entities are simply classified as "abstract " (e.g., "art", "the law").

Regarding the second criterion - some expressions can refer either to an abstract or a concrete object. For example, *newspaper* can be used to refer to a physical object made of paper (*[this newspaper] is very heavy*), to its content (*[this newspaper] is very unreliable*) or to the newspaper publisher (*[The newspaper] hired a new CEO*. The strategy adopted here is to introduce a few categories for entities whose mentions are often hard to categorize, so that you don't have to decide (e.g., "disease"). Else, we use the category "undersp-onto" for the first mention of an entity in case context doesn't make it clear if the mention is used in a concrete or abstract way. (Subsequent references may clarify this.)

[That newspaper]$_{undersp-onto}$ is awful. There isn't a single interesting story in [it]$_{abstract}$.
[That newspaper]$_{undersp-onto}$ is awful. [It]$_{organization}$ fired all its journalists.

Attribute Values

You should consider these values *in the order given*: e.g., consider if an entity could be characterized as an action or event before considering if it could be classified as abstract in a more general sense.

§55 "person"

This value should be used for all NPs that refer to people:

> [You]$_{person}$ should tell immediately to [[your]$_{person}$ doctor]$_{person}$
>
> Does [any of [[your]$_{person}$ friends]$_{person}$]$_{person}$ know [this]?
>
> [[This table's] unusual materials and coloring] allow [scholars]$_{person}$ to link [it] to [a written source] and [a particular building]

§56 "animate"

This value should be used for animate entities other than human beings, such as animals, gods and other supernatural beings, or humanoid robots, as well as for any entity that in your view is displaying agency in the sentence where it is mentioned.

> [Pigs]$_{animate}$ are really very clean.
> [The gods]$_{animate}$ are smiling down on us today.

Note however that organisations have their own category so they should not be marked as "animate" even in sentences where they are attributed agency:

> [Microsoft]$_{organisation}$ increased its annual profit.

§57 "substance"

This value should be used for any NP that refers to substances such as water or gold, but also orange juice in TRAINS.

> [[This table's] marquetry of [[ivory]$_{substance}$ and [horn]$_{substance}$]]
> M: When [you] get [there], fill [it] with [orange juice]$_{substance}$
> [Estracombi TTS patches] contain [[oestradiol]$_{substance}$ and [norethisterone acetate]$_{substance}$]

§58 "medicine"

Another subtype of concrete objects that we want to treat separately are medicines:

> [Estracombi TTS patches]$_{medicine}$ contain [[oestradiol] and [norethisterone acetate]]
> Read [this leaflet] carefully before [you] start using [[your] medicine]$_{medicine}$

§59 "concrete"

This value should be used for references to any other entity that can be touched–e.g., train components or fruit in TRAINS, or all art objects described in the museum domain.

> M: [I] want [you] to take [a boxcar]$_{concrete}$ from [Elmira] and load [it]$_{concrete}$ with [oranges]$_{concrete}$
> [[This table's]$_{concrete}$ unusual [[materials] and [coloring]]] allow [scholars] to link [it]$_{concrete}$ to [[a written source]$_{undersp-onto}$ and [a particular building]$_{concrete}$]

Notice that in this example *a written source* is underspecified between a physical entity (e.g., a manuscript) and an abstract one (the manuscript's informational content), so it is classified as "undersp-onto".

In the pharmaceutical domain, symptoms of diseases should be treated as concrete when they are physical entities that can be touched (hot sweat, red marks), but other symptoms such as a cough are events so should be annotated as "plan"s.

§60 "organization"

This value should be used for all references to organizations, whether expressing agency or not.

> [[Ideal's]$_{organization}$ directors] rejected [that offer].
> Under [the agreement], [HOFI]$_{organization}$ will own [87.2% of [the combined company]$_{organization}$].

§61 "space"

This value should be used for references to geographical entities that clearly occupy a space such as rivers, mountains, streets, and squares. It should also be used for references to countries or towns when clearly used in a geographical sense, such as the towns in the TRAINS world (as opposed to metonymical references to organizations running that country or town, which should be marked as organization).

> [Lake Maracaibo]$_{space}$ is in [Venezuela]$_{space}$
> Let's send [the engine] to [Dansville]$_{space}$
> [Venezuela]$_{organization}$ had a deficit of [$1 billion] [last year]

§62 "disease"

We use this value for all references to diseases such as breast cancer and epilepsy, both when used generically and when used to describe a specific event of catching a particular disease.

> [Multiple sclerosis (MS)]$_{disease}$ is [a disease of [the central nervous system]]
> [You] should not take [this medicine] if you have [a medical condition]$_{disease}$ which makes [you] more at risk of developing [blood clots].

Notice that in the example just shown, *blood clots* is not marked as "disease", but as a "concrete" object. We do this for every term indicating a medical problem which a body part, as in *a broken leg* or *damaged lung*. We also treat every event with medical implications, such as *heart attack* or *stroke*, as an event rather than a disease.

§63 "time"

Every entity that doesn't fit in the categories above is considered a sort of abstract entity, but we recognize a few different types of abstract entities. The first example are temporal expressions. We include in this category both dates such as *1988* (also implicit, like *the century*) and relative expressions such as *last year*. We also mark as "time" expressions referring to durations, such as *three years*.

> [The Thornton mall] opened [Sept. 19]$_{time}$.
> [Sanford Sigoloff, [chief executive of [L.J. Hooker]]], said [yesterday]$_{time}$ in [a statement] ...
> [The workers] are disgusted at being asked to work ''many hours''.

§64 "numerical"

Most numerical expressions in our corpora are used predicatively, as values of attributes. These uses are not marked as referring in ARRAU and therefore their category is not annotated.

> [Canadian Utilities] had [[1988] revenue of [C$1.16 billion]]
> [Enron] is considering building [gas-fired power plants] ... at [a cost of [about $300 million]] ...

The "numerical" value should only be used for numerical expression *not* used predicatively.

> [[Britain's] government] plans to raise about [£20 billion ([$31.05 billion])]$_{numerical}$from [the sale of [most of [[its] giant [water] and electric utilities]]]

§65 "plan"

Actions, events, and event types such as plans should be classified as belonging to the special category "plan".[2] We consider an 'event' every entity which is not concrete according to the test before, does not refer directly to a span of time, and yet has a duration. Examples of events are wars and battles, social happenings such as dinners, elections, or marriages, and medical situations other than diseases,

In [the Dutch wars of [1672] - [1678]]$_{plan}$

Then [we] went to [dinner]$_{plan}$

[Bill] had [a heart attack]$_{plan}$

Examples of actions are the parts of a plan in the TRAINS domain, which can be referred to deictically:

M: then let's send [Engine E1] to [Avon]

[that]$_{plan}$ should take about two hours.

All gerunds should also be classified as events.

[Smoking]$_{plan}$ damages [[your] health]

§66 "abstract"

This value should be used for all other non-concrete objects, from references to abstract concepts such as *art*, *law*, *alchemy*, *ideas*, to references to properties of entities such as *height*, *weight*, etc.

[One stand] was adapted in [the late 1700s or early 1800s century] to make [it] [the same height as [the other]]$_{abstract}$

References to abstract events such as 'life' should be treated as "abstract".

§67 "undersp-onto"

This value should be assigned to all NPs that could be classified as either abstract or concrete. We already discussed the use of "undersp-onto" as the categorization of the first mention of a polysemous word like *newspaper* that could refer either to the physical object or to its content. Another such case are mentions of films or books:

[John] read [War and Peace]$_{undersp-onto}$

[I] only saw [Memento]$_{undersp-onto}$ for [the first time] [last year]

[2] We originally started using the name "plan" as most eventualities in TRAINS were plans, then we kept the name for backward compatibility.

Possible Difficulties

- In the case of pronouns, complementizers, and other headless anaphoric expressions, use the ***Category*** value of the antecedent.

- NPs with a head should generally be classified according to the type of object denoted by the head. One exception are measure-NPs: in the case of an NP such as [4 mgs of [oestradiol]] it is the substance actually measured that should be looked at when deciding how to classify; so this mention should be classified as "substance", like [oestradiol], rather than "numerical" because of milligrams.

- In the case of free relatives, the category depends on the category of the trace:

  `[what [you] need to know about [Nerisone]`$_{medicine}$`]`$_{abstract}$
  `[what's in [[your] medicine]`$_{medicine}$`]`$_{substance}$

- Collective entities should be classified on the basis of their parts: so, [a group of people] should be classified as "person", not as "abstract" because it has people as 'parts'. By the same reasoning, [this pair of [coffers]] should be classified as "concrete" since it was two concrete objects as parts.

- Generic references to types should be marked according to the value that would be given to their instances: e.g., the generic reference to dinosaurs in the sentence *[Dinousaurs] went extinct at the end of the Cretaceous* should be marked as "concrete" even if the type is abstract. NPs such as [a type of ...] should be classified according to the type of object: so [a type of oestrogen] should be classified as "substance", whereas [a type of art] should be classified as "abstract".

Genericity

The ***Generic*** attribute is used in the ARRAU corpus to annotate the 'genericity status' of an NP–a property meant to encode in a compact way information about both ***genericity*** and ***scoping*** ('unselective binding' in DRT terminology).

Genericity is a complex semantic property. Some cases are quite clear. A clear case of generic NP are bare plurals that refer to ***kinds*** in the sense of Carlson (1977)–i.e., types of entities a property of which is being described, as in the following example:

(1) [Cars] are wheeled motor vehicles used to transport passengers.

(the value "generic-yes" of the ***generic*** attribute is to be used for such cases). Conversely, there are NPs whose interpretation is clearly ***episodic***, i.e., non-generic: they refer to specific entities, like the marked NPs in the following example from the RST section of the ARRAU corpus:

```
[Mr.  Uhr] said [Mr.  Petrie] or [his company] have been
accumulating [Deb Shops] stock for several years, each
time delivering a similar regulatory statement.
```

Unfortunately however things are not always so simple. Bare plurals can also be used non-generically, as in the following example:

```
I found [moths] in the wardrobe.
```

And other types of NPs besides bare plurals, such as indefinite and definite NPs, can also be used to express in certain types of generic sentences:

(2) a. [A car] is a wheeled motor vehicle used to transport passengers.
    b. [The car] is a wheeled motor vehicle used to transport passengers.

But not in others:

(3) a. [The Trylobite] went extinct around 250 million years ago.
    b. ??[A Trylobite] went extinct around 250 million years ago.
    c. [The Trylobites] went extinct around 250 million years ago.
    d. ??[The cars] are wheeled motor vehicles used to transport passengers.

The coding scheme used in ARRAU is based very loosely on the theory of genericity originally proposed by Carlson and further developed in ***The Generics Book***. According to this theory there are two ways in which an NP can express 'genericity'. In examples like (1) or (3a), the NP is used as the proper name of a kind ('cars', 'trylobites'). But always according to this theory, NPs can also become generic by being bound by an implicit 'genericity operator'. So in cases like (2a), the indefinite NP *a car* is not intrinsically generic, but gets its genericity through being bound by an implicit generic operator: i.e., the sentence means something like 'Generically speaking, cars are wheeled motor vehicles ...'. We will ***not*** ask you to distinguish between the two ways in which an NP ends up being interpreted generically: in both cases, you are just asked to mark the NP as "generic-yes". But we do want to differentiate the cases in which an NP is in the scope of an implicit generic operator from the cases in which it expresses, or is in the scope of, another ***explicit*** operators such as a nominal quantifier ("operator-iquant"), indicated in bold in the following example:

```
[Most cars] are wheeled motor vehicles used to transport
passengers.  (``operator-iquant'')
```

a temporal adverbial:

```
Often, [cars] are wheeled motor vehicles used to transport
passengers.  (``operator-tquant'')
```

or a modal, including negation which we consider a form of modal operator:

```
[Cars] could be wheeled motor vehicles used to transport
passengers.  (``operator-modal'')
I do not have a [car].  (``operator-modal'')
```

So you are asked to start by considering whether the markable is in the scope of an explicit operator (quantifier, modal, question, or imperative).

Another complication in the scheme is due the fact that in many cases of use especially of bare singulars referring to substances such as *gold* or *water* it is difficult to decide whether the bare singular is used as the name of the stuff or to refer to a specific amount.

```
I found [gold] in the hills.
```

So we will not ask you to make these distinctions; instead, we introduced a number of ***underspecified*** values for these cases, such as "undersp-substance", that you should use in the cases Other cases in which you should not attempt to decide whether the reference is generic or not include references to objects such as books or products that could be interpreted either as referring to the product in general or to a specific instance of the product, for which you should use the value "undersp-replicable"

```
I am reading [War and Peace].
```

Genericity values and how to specify them

In order to decide about the value for the ***Generic*** attribute for a markable, you should consider the following categories ***in the order given***–i.e., earlier categories have precedence over the later ones.

§68 No-generic

Your first task is to exclude markables that cannot be generic. This includes in particular markables that are themselves quantifiers, as in the following example, where *most women* is a quantifier. Such markables should be marked as "no-generic", as in the following example from the GNOME subset of the ARRAU corpus.

```
[most women] have their ears pierced.
```

(Note: in the current version of the ARRAU coding scheme you are required to mark these markables as having a "quantifier" value for their ***Reference*** attribute, which means that you should not be asked to mark the ***Generic*** attribute for such markable–but if this happens use "no-generic".)

§69 Explicit Operators

If "no-generic" does not apply, you should next determine whether the markable may be in the scope of an explicit operator. Operators that can bind markables include, first of all, what we call in ARRAU ***temporal quantifiers***, i.e., temporal adverbials such as *often* or *always*. In the following example from the RST subset, the singled-out markables *investors* and following pronominal references to the same entity are references to an entity that is bound by the temporal adverbial *frequently* and should therefore be marked as "operator-tquant".

```
in the relatively unregulated Indian stock market,
[investors] frequently don't know what [they] are
getting when [they] subscribe to an issue.
```

A second case of explicit operator are nominal quantifiers, that we call ***individual quantifiers*** (i-quant). In the following example, the markable *a similar regulatory statement* is in the scope of the quantifier *each time* so should be marked as "operator-iquant".

```
Mr. Uhr said Mr. Petrie or his company have been
accumulating Deb Shops stock for several years, each
time delivering [a similar regulatory statement].
```

(Note that there is a complication in this case: the quantifier *each time* is in fact a quantifier over time periods. We follow however the convention of always using "operator-iquant" for nominal quantifiers, only using "operator-tquant" when the quantifier is a temporal adverbial.)

Conditionals such as *if*-constructions generally contain an implicit quantifier. Markables referring to entities bound by a conditional as in the following example should be given a value of "operator-conditional".

```
if the economy slips into [a recession], then this is
not a level that is going to hold.
```

Modals are the next class of operators able to scope over noun phrases. Markables bound by a modal operator such as the auxiliary modal *should* in the following example from the RST subset should be marked as "operator-modal".

```
Mr. Gandhi said industry should build [plants] on the
same scale as those outside India ...
```

The value "operator-modal" should also be used with markables in the scope of modal verbs such as *need* in the following example from the TRAINS subset.

```
u : there 's an OJ factory
s : you need [oranges] to make [orange juice]
```

And for markables in the scope of negation:

```
s : don't make [orange juice]
```

Questions can act as scopal binders as well. In the following example, all identified markables are in the scope of the question and should be marked as "operator-question".

```
Who can make [the better decision], [the guy who has
[10 seconds to decide [what to do]]] or [the guy who
has all the time in the world]?
```

Finally, markables can be in the scope of an ***imperative***, as in the following example from the TRAINS subset of the ARRAU corpus. You should use the value "operator-instruction" for these cases.

```
take [a boxcar] from [Elmira] and load [it] with
[oranges]
```

§70 Underspecified Values

If the markable is not in the scope of an explicit operator as in the example above you should next check if it belongs to one of the categories for which no decision about genericity should be made. At present we are doing this for three types of references.

The first class of markables for which you should leave the genericity judgment underspecified are references to ***substances*** like the references to *ivory* and *horn* in the following example.

```
This table's marquetry of [ivory] and [horn], painted
blue underneath, would have followed the house's
color scheme.
```

The second category that you should leave underspecified are references to objects like books or products that are ***replicable*** in Benjamin's sense– i.e., of which there may be multiple copies, and furthermore, for which it is not clear whether the reference is to the content itself or to the physical objects. You should use the value "undersp-replicable" for these markables. For instance, all the identified markables in the following example should be marked as "undersp-replicable".

```
Texas Instruments Inc., once a pioneer in portable
computer technology, today will make a bid to
reassert itself in that business by unveiling [three
small personal computers].
The announcements are scheduled to be made in Temple,
Texas, and include [a so-called "notebook" PC that
weighs less than seven pounds, has a built-in hard
disk drive and is powered by [Intel Corp.  's 286
microprocessor]].
```

The last category of markables for which we always leave genericity judgments underspecified are references to ***diseases*** like *influenza* or *measles*, as in the following example.

```
Most children at my son's nursery caught [chickenpox]
this past Winter.
```

§71 Generic-Yes and Generic-No

If none of the categories above apply, your markable belongs to one of the categories for which we do ask you to make a genericity judgment.

Intuitively, the distinction you have to make is between references to types ("generic-yes") and instances of these types ("generic-no"). In general, names of objects other than the objects in the categories above–e.g., references to people, organizations, locations–should be marked as "generic-no", whereas references using bare plurals to types of objects should be marked as "generic-yes". In the following example, the identified markables should be marked as "generic-no":

```
While [Compaq] sells [its] machines to businesses
through computer retailers, [Texas Instruments] will
be selling most of [its] machines to the industrial
market and to value-added resellers and
original-equipment manufacturers.
```

Whereas the following markables should be marked as "generic-yes":

```
While Compaq sells its machines to [businesses]
through [computer retailers], Texas Instruments will
be selling most of its machines to the industrial
market and to [value-added resellers] and
[original-equipment manufacturers].
```

One type of markables that are generally marked as "generic-yes" are nominal modifiers such as *original equipment* in the example we have just seen:

```
While Compaq sells its machines to businesses through
computer retailers, Texas Instruments will be selling
most of its machines to the industrial market and to
value-added resellers and [original-equipment]
manufacturers.
```

or *oj* in the following example from TRAINS

```
u :  there 's an [oj] factory
s :  you need oranges to make orange juice
```

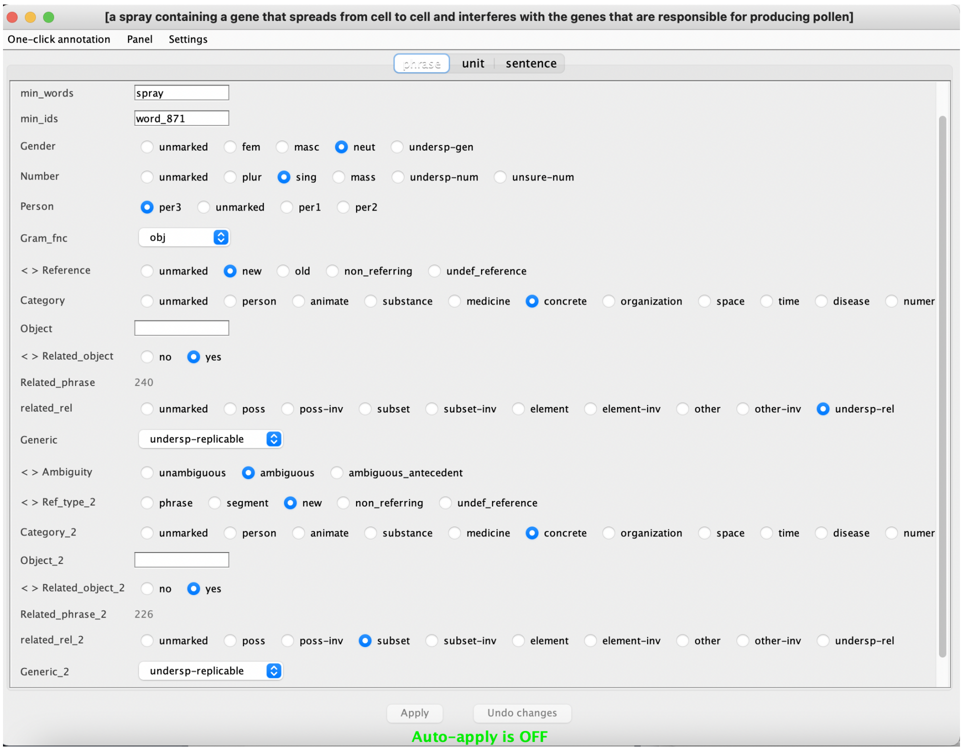
or *time* in the following example from TRAINS:

```
is now midnight and the shipload must be unloaded by
one p.m.  so we've got some serious [time]
restrictions ...
```

Note that all generic references to the same kind / type should be marked as coreferent!!

§72 Undersp-generic

You should use the value "undersp-generic" if the markable clearly belongs to either the category "generic-yes" or to the category "generic-no" but you're not sure which is the appropriate value.



[a spray containing a gene that spreads from cell to cell and interferes with the genes that are responsible for producing pollen]

min_words: spray
min_ids: word_871
Gender: ○ unmarked ○ fem ○ masc ● neut ○ undersp-gen
Number: ○ unmarked ○ plur ● sing ○ mass ○ undersp-num ○ unsure-num
Person: ● per3 ○ unmarked ○ per1 ○ per2
Gram_fnc: obj
< > Reference: ○ unmarked ● new ○ old ○ non_referring ○ undef_reference
Category: ○ unmarked ○ person ○ animate ○ substance ○ medicine ● concrete ○ organization ○ space ○ time ○ disease ○ numer
Object:
< > Related_object: ○ no ● yes
Related_phrase: 240
related_rel: ○ unmarked ○ poss ○ poss-inv ○ subset ○ subset-inv ○ element ○ element-inv ○ other ○ other-inv ● undersp-rel
Generic: undersp-replicable
< > Ambiguity: ○ unambiguous ● ambiguous ○ ambiguous_antecedent
< > Ref_type_2: ○ phrase ○ segment ● new ○ non_referring ○ undef_reference
Category_2: ○ unmarked ○ person ○ animate ○ substance ○ medicine ● concrete ○ organization ○ space ○ time ○ disease ○ numer
Object_2:
< > Related_object_2: ○ no ● yes
Related_phrase_2: 226
related_rel_2: ○ unmarked ○ poss ○ poss-inv ● subset ○ subset-inv ○ element ○ element-inv ○ other ○ other-inv ○ undersp-rel
Generic_2: undersp-replicable

Figure 9: The duplicate attributes that appear in the MMAX window when ***Ambiguity*** is set to "ambiguous"

§73 Other Values

For the moment you should not use the values "episodic-no". Use "unsure-generic" if you really do not know what to do in a particular example.

Ambiguity

General principles

An ***ambiguous*** markable is one which has two (or more) ***alternative*** interpretations. For example, the markable `it` in the following snippet may refer either to the engine or to the boxcar, but not to both, although we can't determine which just on the basis of the text.

```
S: Be careful hooking up [the engine] to [the boxcar]
because [it] is faulty
```

An ambiguous item should have ***separate*** markings for each interpretation: in the example above, `it` should have a "phrase" reference pointing to *`the engine`*, and a separate "phrase" reference pointing to *`the boxcar`*.

To specify a second reference, select the value "ambiguous" for the final attribute, ***Ambiguity***. This will bring up a second set of attributes which you can use as before to mark the second meaning, as illustrated in Figure 9.

Notice that the alternative interpretations may be of the same reference type (phrase / segment) or of different reference types: i.e., a markable may be interpreted as referring either to a phrase or to a segment. You should mark each reference type separately.

How to mark ambiguity

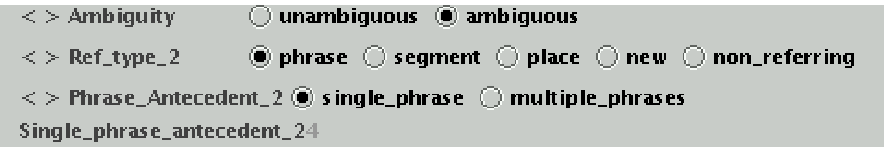
§74 Note that if you decide that a markable is ambiguous between two interpretations, you need to specify all semantic properties of the second interpretation, not just ***Reference*** and antecedent(s). I.e., you will need to specify the ***Category*** of the markable in the second interpretation, its ***Genericity***, etc. This is because often the ambiguity is caused by the fact that a markable could be interpreted generically or not.

§75 A markable may also be ambiguous between an "old" and a "new" interpretation: i.e., whether it refers to an object already mentioned or to a new object.

§76 Markables may even be ambiguous between and "old" and a "non_referring" interpretation. For example, in the following fragment it is not clear whether the *it* refers to a segment (the action of getting the boxcar to Elmira, say) or does not refer at all.

```
M: Let's take engine E1 to get the boxcar to Elmira.
S: I don't know how long [it] will take.
```

In order to mark these ambiguities, choose "old" as the value of the ***Reference*** attribute, and then use the second reference type attribute to specify the values "new" or "non_referring".



< > Ambiguity ○ unambiguous ● ambiguous
< > Ref_type_2 ● phrase ○ segment ○ place ○ new ○ non_referring
< > Phrase_Antecedent_2 ● single_phrase ○ multiple_phrases
Single_phrase_antecedent_24

§77 Only mark an item as "ambiguous" if the two interpretations are distinct, for instance if they refer to distinct objects, or if one interpretation refers to an object and the other does not. If two possible antecedents refer to the same object, just mark the most recent one.

§78 If an item has more than two possible interpretations, only indicate the two most likely ones.

§79 Sometimes, the ambiguity is temporary, but is resolved a few utterances later. For example, the pronoun *it* in utterance 3 might refer to the boxcar or the tanker, but utterance 5 makes it clear that the intention was the tanker.

```
1.  M:  I want you to hook up the boxcar
2.   :  and the tanker
3.   :  and take [it] to Dansville
4.  S:  Okay
5.  M:  When you get there, fill [it] with orange
        juice
```

In these cases, annotate the first markable as ambiguous. The second markable is, of course, unambiguous.

§80 In other cases, you'll find an 'ambiguity chain': a series of anaphoric expressions each 'ambiguous in the same way'. For instance, in the following fragment, the first *it* (in utterance 4.) may refer to either engine E1 or the boxcar; the same is true of the second *it* (in utterance 6).

```
1.  M:  good
2.   :  uhh
3.   :  can we please send engine E1 over to Dansville
        to pick up a boxcar
4.   :  and then send [it] right back to Avon
5.  S:  okay
6.   :  [it] 'll get back to Avon at 6
```

In this case, specify just one antecedent for the second *it* (that is, a pointer to the previous *it*), and use the value "ambiguous_antecedent" for the ***Ambiguity*** attribute.

Correcting errors

§81 If you make an error, you can correct it as follows:

to change the value of an attribute simply select a new value by clicking on the appropriate radio button in the attribute window; the old value will be removed.

to remove the reference of a phrase or segment markable follow the same procedure as for marking a reference, right-clicking on the brackets of the markable you want to remove; the text "Remove reference to this phrase/segment" will appear, and you can click that to remove the reference.

Be sure not to remove all the references of a phrase or segment markable! Also note that if you change the attribute value of a phrase or segment markable and then change it back, all the references you have marked will disappear and you will have to mark them again.