Annotation scheme for information status for ISNotes 1.0

Katja Markert, School of Computing, University of Leeds

1 Introduction and Summary

This scheme describes annotation for noun phrases according to information status. First, all noun phrases are distinguished as either a mention (referring to an entity) or a non-mention (expletive and pleonastic *it, there*, idioms, reflexives as emphasisers and proper names embedded in other proper names). Then all mentions receive an information status, which is one of old, mediated or new. `old` entities are identical with a previous entity, set of entities, action, sentence or speech act in the text. In addition, generic pronouns as well as pronouns referring to conversation participants are also old. `mediated` entities encompass a variety of entities — they all have in common that they have not been mentioned previously in the discourse (not `old`) but are also not *autonomous*, i.e. they can only be understood by reference to another already mentioned entity or your prior world knowledge. Thus, `mediated` encompasses entities linked to an old or other mediated entity explicitly (`syntactic, aggregate` or `func`) or implicitly (`bridging, comparative`). It also encompasses entities whose uniqueness is determined by world knowledge or the wider discourse context. All other entities are `new`.

2 Mentions and Non-Mentions

Our definition of non-mentions follows the OntoNotes 4.0 annotation scheme. All mentions are then further annotated for IS.

The following are non-mentions. All other marked up items are to be annotated for information status. Non-mentions are either expletive or pleonastic *it/there*, reflexive pronouns that are used as emphasisers, parts of idioms or parts of proper names.

2.1 Expletive and pleonastic *it* and existential *there*

There are occurrences of *it* and *there* that do not refer to any specific entities but are syntactically necessary. These are to be annotated as `nonmentions`. Examples are:

1. ***It*** *is raining.* (You cannot ask "who or what is raining?")
2. ***It*** *is a shame that I did not meet you yesterday.*
3. ***There*** *was a woman I loved.*
4. ***It*** *is good that you came.*
5. ***It*** *was Maggie Thatcher who destroyed the unions.*
6. *I know that* ***it*** *is 5 o'clock.*

Many of these cases already appear in the tool as (it)*EXP* but some don't.

2.2 Idioms

Noun phrases that are parts of idioms are not to be annotated.

1. *If you can make* ***it*** *there, you'll make* ***it*** *anywhere.*
2. *raining* ***cats*** *and* ***dogs***
3. *In* ***fact***
4. *For* ***instance***
5. ***You*** *know* (if idiomatically used)
6. *They hit* ***it*** *off.*

2.3 Nested Proper Names

We see proper names as atomic, i..e proper names nested in other proper names are not mentions. Examples are

1. *University of* ***New York*** ("New York" is not a mention as part of the bigger named entity "University of New York")
2. ***Chicago*** *Board of Trade* (Both Chicaco and Trade are not mentions)

Note that this is only true for proper names. In other cases, all nested noun phrases are mentions:

1. The example [window of [the house]] has two mentions
2. The example [[Chicago] wheather] has two mentions

2.4 X itself

Most reflexive pronouns (*itself, themselves, himself* etc.) are premarked as `old` as they refer back to a previous entity such as in *He washed **himself***. However, in rare cases they are only used as an emphasiser in the construction *x itself/herself/...*. An example in context is *George Bush **himself** did believe that Iraq had WMD*. Here, *himself* is just an emphasiser and should be marked as a non-mention. This is the only grammatical construction with reflexives to be considered a non-mention.

3 The main categories: old, new and mediated

Old entities are entities that have been mentioned before within the discourse and are coreferent to a previous concrete or abstract entity or are discourse deixis (generic pronouns, for example). In contrast, ew entities and mediated entities have not been mentioned before. The main distinction between mediated and new entities is that mediated entities are *not autonomous*, i.e for their understanding it is necessary to either link them to a previous discourse entity or to your prior world knowledge (which might be independent of the text). New entities are autonomous, i.e. introduced without necessitating a link to the discourse or your world knowledge. It is important to realise that world knowledge does not mean lexical knowledge. We assume that you do know in any case what words mean. World knowledge means linkage to facts about or entities in the real world —this knowledge helps to interpret mentions which you could **not** interpret without this knowledge.

4 Old

`Old` entities are coreferent with previous entities or generic pronouns or pronouns referring to speakers in a conversation or quote. If they are coreferent with a previous entity, the previous entity is called an *antecedent* — in the following examples, antecedents are underlined. You do not have to mark the antecedent in the annotation tool.

4.1 Coref: Premarked: not allowed for annotation usage.

An entity that is identical to a before-mentioned entity. The coreference annotation has been imported from the OntoNotes coreference annotation and is to be unchanged.

These are already premarked in the annotation tool, including the antecedent. Please do not annotate more such coreference links.

4.2 other_coref

There are some rare cases where coreference is not premarked. Therefore, we add some coreference annotation to the OntoNotes coreference annotation but mark these with a separate tag so that the two layers are distinguishable.

There are three such cases which you should then mark as `other-coreference`. IN these cases, again the entities are identical.

- A mention refers back to something other than a noun phrase, for example refers to a whole sentence or remark. This is especially frequent with the pronouns *this/that* on their won (see second example below). Examples:

  1. *He wonders whether will ever complete his PhD. **This question** occupies him daily.* (Here, this question does not corefer with "PhD" but to a whole string "whether he will ever complete his PhD".)
  2. *John never contacts his supervisor. **That** is unacceptable.*

- Coreference between a singular and a plural entity. This you can mark as other_coref **if and only if** the coreference is crystal clear (very, very rare).

  1. *They made memory chips. To do this, they placed the **chip** ...*

- A plural phrase that is the aggregate of two or more previous phrases. [In less clear cases, this is more probably `bridging/set`.]

  1. *John was hired as an accountant. Then Keeran was hired as an accountant. **The two accountants** did not get along.*

It is essential to note that you are not supposed to improve the existing coreference annotation, no matter how much you disagree with it. In particular, the following are allowed or not allowed:

- ↯Additional coreference links between two already bracketed noun phrases (not proper names) with the same number (both singular or both plural).

  Examples: *School boys . . .* the *the 16-year old school boys*.

  The government . . . the administration

  We start with the assumption that standard coreference between two same-number mentions is correctly pre-annotated via the OntoNotes corpus.

- ✓Additional coreference links when a pronoun has not been annotated.

- ✓Additional coreference links between two identical proper names.

For IS corpus users: Please note that the second allowance has led to binding phenomena being annotated as old. This will be corrected in future releases.

4.3 Generic/general pronouns

Definition.

- *Type I: Personal pronouns "you", "we" and "they" that are not used referentially as in* **You** must save for your pension *when talking about everybody, not one specific person.*

- *Type II: The indefinite pronouns "one, anyone,everyone,everybody . . . " when used non referentially:* **Anyone** must save for their pension.

- *Type III: Pronouns "I/me/myself" and "You/your/yourself" when referring to speakers in a conversation or quote.*

4.4 Relative

Relative pronouns should be annotated as "relative". An effort has been made to exclude these in advance from annotation so you should not have to use this category.

1. *I met a man* ***who*** *wore red trousers.*

5 Mediated

Mediated entities not mentioned in the discourse previously, but they are also not *autonomous*, i.e. not fully understandable without prior context or prior world knowledge. They need to be one of the following:

1. Indicated by lexical markers (*such, similar, other, comparative adjectives*) to be comparative to other previously mentioned entities or real-world entities.

2. Understandable because they are linked by a limited number of syntactic constructions to a previously introduced or other mediated entity. The constructions are limited to possessives, the of-genitive, any prepositional phrases and proper name premodification. No other constructions are allowed to be considered for syntactic mediation.

3. Aggregated via a coordination using at least one mediated or old entity as one of its components.

4. Function entities referring to items on a scale.

5. Understandable only by making an act of inference to a previous entity, which is not identical or near-identical (bridging).

6. Understandable only by reference to prior world knowledge of the typical reader. This is only applicable to mentions starting with "the/this/that" or proper names.

In cases of conflict where you think more than one of the above subcategories applies, they have preference in the order given, replicated by left-to-right order in the annotation tool

5.1 Comparative

Comparative phrases include a premodifier that makes clear that the entity is compared to a previous one. The two entities are not identical in the real world but their types are the same. Thus, in *George Bush — other presidents*, the *other presidents* are not the same as George Bush but of the same type (presidents). The mention would not be understandable on its own — you could not go home and start a sentence without context saying *Other presidents ...* without inciting the comment *What do you mean: other presidents than whom?*.

The entity compared to can be both within the text or outside the text. If within the text, the entity compared to is underlined in the following examples.

- Set complements, normally introduced by *other, another, a different, (...) else* and so on.

    1. *He hurt his right hand. Thankfully,* ***his other hand*** *is fine.* (Here, *his other hand* is a complement to *his right hand*. Note that comparative

takes precedence over `syntactic` although a possessive *his* is mentioned.)

2. *FED wants to raise interest rates to combat inflation.* ***A different solution to inflation*** ... (Here, we refer to a *solution different to raising interest rates*)
3. *I don't really like my job.* ***Everything else*** *is going well, however.* (Everything other than my job).
4. *I like* ***other films than westerns***. (In this case, the *other films than westerns* are again a comparative entity with the entity *westerns* being the one compared to. In these cases, the two entities are nested into each other.)
5. *I like pork, veal and* ***other meat products***. (Make sure to create a new entity consisting of *pork,veal* as antecedent in such cases.)
6. **Counter Example:** *I like him. On* ***the other hand****, he can be annoying.* (*other* can be part of idioms —then it is not a mention).

- Apart from contrasting sets, one can also emphasise similarity. This is often done by *such, similar, same* and other such words.

  1. *The FED wants to raise interest rates to combat inflation.* ***Such solutions*** ... (here we talk about solutions similar to raising interest rates.)
  2. *I bought a blue dress. Then I saw my neighbour in* ***a similar dress***.
  3. *John bought a car. Then, Jack went out and bought* ***the same car.*** (here, it is not really the same physical car they bought but the same model. Thus not coreference but comparison).
  4. **Counterexample:** *John is* ***such a loser***. (*Such* is a highly ambiguous words. It can be part of idioms. Here it is just an emphasis of how big a loser John is and therefore not comparing the "loser" mention to anybody else. It should be annotated as *new* here.)

- Comparative adjectives such as *smaller, less interesting* can all indicate comparative mediated mentions. This applied to comparative adjectives only, not to superlatves (*most, smallest, most interesting*) or the base form of adjectives (*many, small, interesting*).

  1. *Celebrities stay in the 5-star luxury ocean view suites.* ***Poorer visitors*** *stay in the 3-star neighbouring hotels.* (Here, we mean *poorer visitors* than celebrities).

2. **Counterexample:** *We treat **older people** badly.* (Without a special context, this does not mean *older people than xyz* but is a more polite term for elderly people and not comparative. It is understandable without sb having to ask *older people than whom*?)
3. **Counterexample:** ***More than 25% of patients** are unhappy with the service they receive.* (Unclear to what this is compared to; don't use comparative).

Comparative entities can also compare an entity to sth not explicitly mentioned in the text:

1. *Ministers have said that the costs of the new university loan system might be prohibitive. This could lead to **fewer university places.*** (This means *fewer university places than currently* which is not explicitly mentioned.)

Rephrasal test. Normally you can rephrase that by one of the constructions "other than/different to", "such as/similar to " or "more/bigger than".

Antecedent linkage. Once you have decided on marking an entity as comparative, you can link to the entity compared to or say that there is no antecedent in the text itself.

Additional bridging links. In very rare cases, there is also a bridging link from a comparative entity. You can give such additional links after linking to the standard comparative antecedent. An example is given below.

1. *Aids has many different symptoms. It can cause dementia. **Other symptoms** include .ldots*. (Here, *other symptoms* is comparative, and means *other symptoms than dementia* so *dementia* is the comparative antecedent to link to. In this particular case, you still miss which disease the symptoms are of, so you should link to *AIDS* as additional bridging link. The complete rephrasal is *other symptoms of AIDS than dementia*).

5.2 Syntactic: PPs, possessives and premodification only

Some entities have not been mentioned before but have been explicitly anchored to an old or already determined mediated entity and are therefore understandable. This could in theory be done via a range of syntactic constructions but we restrict ourselves to the following:

- Possessive pronouns or Saxon genitives as the anchor. If the possessive pronoun or Saxon genitive possessor (X in X's Y) is old or mediated, then the whole phrase is mediated.

1. *John has problems with his family.* ***[[His] father]*** *is a an alcoholic.* (Here *his father* is mediated as *his* is old.)
2. *John has problems with his family. I, however, really liked* [[John's]father]. (Here, again *Johns father* is anchored by the old entity *John* and is therefore mediated.)
3. *John has problems with his family.* ***[[[His]father's]*** *alcoholism] ...* (Here, *his* is old, therefore *his father's* is mediated/syntactic and therefore *his father's alcoholism* is mediated/syntactic as well. Please note the precedence of mediated syntactic over the set constructions that would allow *his father's alcoholism* to be interpreted as a set member (bridging) of the aforementioned problems.)

- Of-genitives.

1. *John has problems with his family.* ***[The alcoholism of [[his]father]]*** *ldots* (Here again, *The alcoholism of his father* is mediated/syntactic.)
2. **Counterexample:** ***[The problems of [all families]]*** *resemble each other.* (Here, *all families* would be new and therefore *The problems of all families* is new.

- Proper name premodifiers.

1. ***[The [Federal Reserve] boss]*** *raised interest rates.* (If the Federal Reserve is judged to be mediated/world knowledge, then the *Federal Reserve boss* is mediated/syntactic. This is therefore judged equivalent to [The boss of [the Federal reserve]])

- Other Prepositional phrases attached to the noun phrase can also mediate.

1. *I bought a new car.* ***[The seats in [the car]]*** *are out of leather.* (As *the car* is old, *The seats in the car* are mediated/syntactic.)
2. *He worked as* ***[professor in [law] at [Cambridge]]***. (Given that *Cambridge* is marked as world-knowledge/mediated, *professor in law at Cambridge* is mediated/syntactic. The bracketing shows that *Cambridge* is attached to *professor* and not to *law*.)
3. **Counterexamples**. We only allow prepositional postmodification to mediate. Therefore, other postmodifications do not allow for a mediated annotation. *John is my best friend ...* ***The house John bought*** — although semantically one could argue that *the house John bought*

is mediated by the old entity *John*, this is not done by a prepositional phrase or Saxon genitive. Therefore we would annotate *the house John Bought* as `new` if not mentioned before. Similarly, appositions are not used to mediate.

4. **Counterexamples:** The prepositional postmodification needs to be to a noun phrase. Therefore, ***The idea of going to London*** is not mediated by *London* as *"of going"* is a verbal phrase. Note here that *to London* is attached to the verb phrase '*going to London* so cannot be used for mediating *idea*. This can easily be tested by trying out *the idea to London*, which does not make sense.
5. **Counterexamples:** You have to make sure that the mediating postmodification applies to the noun phrase in question (as shown by the square brackets). Thus, in *[I] met [a man] in [Leeds]*, *Leeds* is mediated/world_knowledge but not a postmodification of *a man*. Therefore *a man* would be `new`.

General guidelines. Annotate these entities/mentions from the inside (most embedded parenthesis first). Consider recursive mediation as in *His father's alcoholism*.

Please note the following:

- ✓The entity that mediates needs to be bracketed inside the entity you annotate.
- ↯Adjective premodification does not mediate. Example: *The American president*. *US* counts as an adjective in premodification. Treat these cases either as new, bridging or worldknowledge/text, as appropriate.

5.3 Aggregation (see Nissim et al)

The subtype aggregation is only used with coordinated NPs. It is used when not all coordinated NPs are new. It is enough that one of the coordinated Nps is old or mediated, for the coordination to be mediated/aggregation.

Typical coordination constructions include *and, or, either ... or, not (only) ... but, neither ... nor*

- ***The president and his daugher*** *go to New York.* Mediated/aggregated as *the president* is mediated/world knowledge.
- ***John and my daughter*** *go to New York.* Mediated/aggregated as *my daughter* is mediated/syntactic.

- ***Not only George Bush but also Barack Obama*** *...*
- **Counterexample:** ***John and Mary*** *go to New York*. New if both John and Mary are new.

Aggregate takes precedence over bridging as indicated in the annotation tool. Thus, in the example *Siemens rearranged its management structure. Peter Hahn was named* ***[[director] and [chief executive]]***, both *director* and *chief executive* are mediated/bridging to *Siemens* but the aggregate *director and chief executive* is `mediated/aggregate`.

5.4 Func

The type `func` is used to indicate the relationship between a function and its values. The function needs to be explicitly mentioned and needs to be able to rise and fall.

1. *The temperature rose to* ***30 degrees*** *before dropping to* ***15 degrees***
2. *IBM shares were down* ***3 points***.
3. *IBM shares stand at* ***7 dollars.***
4. *IBM shares went down* ***3%*** *to* ***5 dollars a share***. Note that both the values as well as point/percentage falls and rises are to be annotated with `func`).

5.5 Bridging

As an introduction, it is necessary to learn that nouns have certain *semantic roles or arguments*. For example, *revenge* has somebody that carries out the revenge, somebody the revenge is carried out on, what the revenge consists of and what the revenge is for. As another example, *goal* has an actor who has the goal and sth that the goal is directed at. The noun *father* has only one main role, namely of whom one is the father. Parts (such as building parts like *roof*) have the whole as a mandatory role. Some nouns need no roles at all as they are complete nouns on their won, such as *car* or other vehicles. All nouns can also have non-mandatory roles, for example *revenge* and most other action-related nouns can have a duration. Such roles are described in the online ressource FrameNet `ttp://framenet.icsi.berkeley.edu/}.Textsnowfillsuc` roles, thus in *John's father*, the role of the offspring is filled syntactically. Similar, in *the Mafia took revenge on John* the role of the revengetaker and the person whom revenge is taken on are filled within the clause. Not all roles are always filled in all texts such as in *John*

became a father where the role of the offspring is not filled but the sentence is still understandable.

Roles can also be filled by inference. Thus, in *John worked very hard during the last few months. As **a result**, he now needs a long holiday.*, the cause of the results (working very hard) is not filled explicitly but must be inferred from the previous sentence. This is what we call *bridging* — the filling of necessary roles via inference.

Definition. *Bridging noun phrases are (non-old) noun phrases where the following holds:*

1. *They are* not autonomous*, i.e. not complete on their own. They miss one or more mandatory roles which are not filled syntactically. The missing mandatory role is often noticed as the noun phrase seems to call for a completion via a prepositional phrase* of/for/by*, such as* result of what? *in the above example.*

2. *The missing mandatory role(s) can be filled by a prior entity (entities) present (normally present in a different clause). This entity is called the antecedent of the bridging entity and is underlined in the examples below.*

3. *The antecedent is **not** identical or near-identical to the bridging entity, which distinguished the phenomenon from coreference. In addition, they are also not of the same type (semantic category) as the bridging entity.*[1]*. This difference in semantic type distinguishes bridging from* `comparative` *as well as from coreference.*

Bridging noun phrases can be of any form (definite, indefinite, singular, plural) but are most frequent in a definite form starting with *the*. However, pronouns and proper names are very, very rarely bridging.

The bridging antecedent can be a noun or a verb phrase and in some cases even a speech act (see examples below).

Rephrasal test. You must be able to rephrase the bridging entity by a complete phrase including the bridging entity and the antecedent.

In the example, *The company wrote out a new job. **Two applicants** were suitable.*, the bridging entity *two applicants* can be rephrased by *Two applicants for the new job* and a new job is the antecedent.

If the antecedent is a noun phrase, rephrasals are restricted to a prepositional phrase or possessive/Saxon genitive. Otherwise, they should only include words

[1] There is one exception to this semantic category rule, set bridging, explained below.

in the bridging entity and its antecedent. Changes in determination for the bridging entity (changing *a* to *the*, for example) should not be necessary. The rephrasal needs to sound natural in the whole sentence. If the antecendet is a verb phrase, other constructions are necessary and one needs to be more flexible with the rephrasal.

Further guidance on specific cases (such as cases where one role is syntactically specified but other mandatory roles are not) are after the example section.

Examples and counterexamples with context. Noun phrase, verb phrase or speech act antecedents Examples and counterexamples are given below:

1. *I bought a car. **The seats** are out of leather.* (Here, *the seats* are not understandable and incomplete on their own without linking them back to *a car*. You could rephrase with *I bought a car. The car's seats are out of leather.* This is the only link that makes sense).

2. *I bought a bycicle. **A tyre** was already flat.* (Rephrasal *A tyre of the bycicle*)

3. *The UK elected a new government. Michael Jones was made **prime minister**).* (Functions such as ministers, treasurers, company presidents should be linked to functions in a company or coun try. Therefore *prime minister* here is to be linked to UK. Rephrasal: *prime minister of the UK*)

4. *Our company wrote out a new job. **Two applicants** were suitable.* (Applicants for the job)

5. *There was a terrible earthquake in Haiti. It is awful to see **the suffering the people are going through**.* (Here, the first thing to note is that this entity (suffering...) is not mediated/syntactic as it is not one of the constructions we allow there, even if you link people to people in Haiti via mediated/bridging. Therefore, *the suffering* is mediated/bridging to the earthquake: suffering caused by/from the earthquake. This example shows that more complex noun phrases can also be bridging.)

6. *I travelled to Edinburgh. **The train** was very full.* (Rephrasal: the train with which I travelled. The first example with a verb phrase as the antecedent)

7. *Why do humans collaborate? **The answer** lies in ...* (The answer to the whole speech act of the previous question. We will just link to the wh-question word.)

8. *Glaxo-Smith-Kline started a new drugs trial. They will pay **participating patients** x dollars.* (Rephrasal: Patients participating in the drugs trial. In cases

where the bridging entity has a premodifier, reodering in the rephrasal can be necessary.)

9. *The cave was inhabited during the ice age.* ***The cave walls*** *are covered with prehistoric paintings.* (Rephrase: The walls of the cave. Note that *cave* in *cave walls* is not a separate mention so this is not a mediated/syntactic item.)

10. *He gave evidence against the Mafia hitman. The Mafia killed him in* ***revenge****.* (Here, *revenge* is *revenge* for giving evidence. You could also link *revenge* to *Mafia* for *revenge by the Mafia* but this is not necessary as the sentence says that explicitly. Similarly *revenge* could be linked to *He* as *revenge on him* but again that is expressed sentence-internally. Using rephrasal tests *The Mafia killed him in revenge for giving evidence.* is plausible whereas the combined rephrasal *The Mafia killed him in the Mafia's revenge for giving evidence* does not.)

11. **Counterexample:** He bought *a car*. (Here, *a car* can be felicitously used without further explanation/inference so no bridging is necessary. The entity is `new`. Rephrasal sounds very implausible: *he bought his car*).

12. **Counterexample:** *The fact that I bought a car* . . . (Here again, this is completely understandable without any bridging and the entity is complete on its own. It is again `new`).

13. **Counterexample:** *IBM fell three dollars* ***a share***. (Rephrasal is very strange *IBM fell three dollars a share of IBM???*).

If you link to a verb phrase (generating a new entity) just mark to the main head verb, not the whole phrase.

Several bridging links from one bridging entity In some cases, several mandatory roles are not filled syntactically but can be filled by inference to two different antecedents. In these cases, instead of choosing one we fill all the roles (as allowed in the annotation tool). A condition is that the overall rephrasal using all antecedents is plausible.

1. *He_1 put_2 all his money into shares.* ***The goal*** *was to save for a house deposit.* (Here, *the goal* is *the goal of putting (his money into shares)*. You could also link *goal* to *he* as it is *his goal*. Two mandatory roles for *goal* are missing syntactically and can be filled by inference. The tool allows you to make both links at the same time which you should do. The overall rephrasal *His goal in putting money into shares* is plausible.)

Typical entities that might need more than one bridging links are abstract entities (such as goals, reasons) as well as nominalizations (nouns derived from verbs such as *revenge*).

Mediated/Syntactic entities with a missing mandatory role. `Mediated/syntactic` has precedence as an annotation category over `bridging`. However, it is possible that one role is filled syntactically but there is still a mandatory role missing which can be filled inferentially. In these cases, after annotating the entity as `mediated/syntactic` you are allowed to mark an additional link (called *other_links* in the tool) to the "bridging" antecedent.

1. *He put all his money into shares.* ***His goal*** *was to save for a house deposit.* (*His goal* is `mediated/syntactic` as one mandatory role is filled via a possessive. However, a second one must be filled by linkage to *putting money into shares*.)

Typical entities that might need a bridging link in addition to their syntactic mediation are abstract entities (such as goals, reasons) as well as nominalizations (nouns derived from verbs such as *revenge*).

Set Bridging Set bridging is different from other bridging in two ways:

- The role used is not mandatory as subsets can be created freely for any entity type (*two cars ... the first car*).
- The semantic types/categories of the bridging entity and the antecedent are the same.

We will use set bridging sparingly to avoid annotating all set relationships in a text. You are only allowed to use set bridging when the bridging entity is not understandable without linking it to the antecedent. The bridging antecedent must be a superset to the bridging entity (never the other way around.) The antecedent is always a noun phrase. Often set bridging will be indicated via specific lexical markers (see the first two examples below) although other examples are possible.

1. *I bought eggs.* ***One*** *was broken.* (Here, *one* is a set bridging to *eggs*—only understandable because *eggs* are mentioned before.)
2. *He had two problems.* ***The first problem*** *is unemployment.* ***The second problem*** *is that his wife has left him.* (Here, both bold-faced entities are bridged back to *two problems* with set bridging. They can also be linked back to *He* which should be done in the first case as not expressed in the sentence itself. Here we have bridging without a core role.)

3. *John and Mary bought two books each. **John's books** were on ancient Greece.* (*John's books* is `mediated/syntactic` but again are only fully understandable by making an additional link to the *two books each* mentioned before.)

All other rules (rephrasal test, mediated/syntactic rules) are the same as for standard bridging.

Bridging types, examples without context Typical (but by no means all) cases of bridging include:

- Physical parts of objects and humans (*car — the engine*).
- Attributes of objects such as colour, emotion, prize, symptoms of a disease ...
- Reasons, goals, objects, plans, results and causes of actions.
- Items marked with *one, the first (out of a set), the second (out of a set), some superlatives etc.*. These cause set bridging.
- Instruments or other roles of an action/verb (*was murdered — the knife*).
- Relations between people and between people and possessions such as *the owner, the father, friends, fans, critics*
- Functions such as titles or functions in a company or a country (*vice-president*). Relation between people and a place (*residents, citizens* ...) or people and an object *viewers, readers* ....
- Actions on sb or sth (*country — sanctions (on the country), the product — pricing (of the product), book — the writing (of the book), steel industry — subsidies (to the steel industry), build new houses — approval (for building)*.
- many others, unfortunately too many to list

Antecedent selection for bridging If you have the choice of several antecedents for a bridging item, please follow the following rules:

1. If they have different mandatory roles, then annotate all of them. The relevant example above was *He put all his money into shares. **The goal** was to save for a house deposit.*. This is possible in the tool which lets you add several bridging links with different roles to different antecedents. You first

select one bridge type (here, event) and link to *put*, then you select the second type *other* and link to *He*.

2. If you have only one role to annotate, but several possible (and non-coreferent) antecedents for that one role, use the one where rephrasal sounds the most natural.

3. If you have one bridge type and one antecedent with several realizations in the text, link to the one closest to the mediated/bridging entity. In the Example *Rolls Royce announces changes to its management structure. John Smith is to join the board.* link the board to *its* (which is coreferent with Rolls-Royce). The annotated coreference chains in the tool will show you the closes realization of an entity.

Relations for bridging *Note: this part of the annotation is not yet part of the current ISNotes release. In the current release only part-of and set bridging have their relation type marked. All other cases have relation type "other"..*

The following relations are allowed for bridging. The relations are seen from the viewpoint of the anaphor but the examples are given in the natural text order, i.e. antecedent first.

- *set-of*: The anaphor is a subset of or particular instance of the antecedent set. An *isa* relationship holds between anaphor and antecedent head. We also include pronoun anaphora here *the eggs — one was broken*.

- *has-set*: The anaphor is a superset of the antecedent or the anaphor is a set containing the antecedent instance. An *isa* relationship holds between antecedent and anaphor head. Pronoun anaphora are included here as well.

- *part-of*: The anaphor is a part of the antecedent, including physical parts and members as well as material. Physical parts can be optional. Examples are *the room — the chandelier*, *the house — the roof*, *forest — trees*. Examples for materials include *handbag — the leather*. In cases of doubt, this has preference over other spatial relations.

- *attribute-of*: Anaphor is an attribute of the antecedent, such as *colour, velocity, beauty, truth*.

- *telic/agentive-of-object* This includes telic/agentive roles of an object antecedent such as *film — the producer*, *book — readers/author*, *oil — producers* as well as generalised possession. The anaphor is normally a person/organisation/GPE and the antecedent is an object. We also include generalised possession/ownership here.

- *spatial-relation:* Anaphor and antecedent stand in a spatial relation (containment, neigbouring etc) but are not in a part-of relation. Can be the same semantic class. Examples :*this car — the next car*, *bottle — the beer*. In cases of doubt this is preferred over temporal relations.

- *temporal relation:* Anaphor and antecedent stand in some form or temporal sequence or happen at same time. Can be the same semantic class. Examples: *Vertigo — his next/following movie*

- *role-of*: The anaphor is a person or suborganisation that fulfils a role in an antecedent organisation or GPE. Examples are *Japan — the government/prime minister/president*, *New York — the city council*. We also include residents and citizens here, although their function is less clear.

- *Relative-of*: relation between two persons *rival, enemy, friend, mother* .... Metaphoric and metonymic usage of such relations (for example, two countries being friends or enemies, being a lover of independent film) are also included.

- *has-semantic-role*: Anaphor is state/event (abstract object) and the antecedent has a role for the event (*the country — sanctions (against that country)*, *film — production (of that film)*). This can be related to *person-to-object*; the difference is in the anaphor being abstract instead of a person/organisation/GPE.

  We include here also abstract anaphora that are the aims/reasons/consequences of an event *bankruptcy — reasons* as reasons need a semantic role that states what they are the reason for.

- *is-semantic-role* Anaphor is an event role and the antecedent is an event/state (abstract object). Examples: *murdered — knife*; *flight — the pilot*, *operation — the patient*. Antecedents that are only metonymically an event are included here as well *Vietnam — the casualties (in the Vietnam war*).

- *other* All instances that do not fit into any of the above categories.

5.5.1 World Knowledge or Situated by Overall Text/Discourse Topic

Definition. *This is used in the following case only (all items below have to hold!):*

1. *Only applicable to proper names (including specific time mentions) and NPs starting with* the/this/these/those/that.*)*

2. *No explicit or inferential link to another entity in the text via syntactic mediation, comparative anaphora, aggregation, bridging, coreference.*

3. *Still understandable by your world knowledge or by the overall discourse topic. Not new to you. Uniquely identifiable.*

Examples:

1. *The Pope* (if mentioned for the first time)
2. *The moon* (again if mentioned for the first time)
3. *That Osama Bin Laden*
4. *Margaret Thatcher*
5. *President Obama*
6. *Rooney is going forward and shoots. But **the referee** blows his whistle.* (Even though the word *football match* is not explicitly mentioned, the referee is uniquely understandable due to the wider context.)
7. *Rooney is going forward and shoots. I wish he had played like that in **the first half**.*
8. *I left in **spring*** (specific time/date)
9. *I left **last May*** (specific time/date)
10. **Counterexample**: *John Hunter, a salesman from Leeds* . . . (this would be `new`)
11. **Counterexample:** I am shocked by *violence* (not a proper name, or definite noun phrase)

You can also use this category for definite NPs that are somehow inferred by the text but no direct bridging or coreference link can be established.

6 New

All other entities are new. This includes entities that are (not a complete list):

- Uniquely identifiable NPs that are fully described by pre- or mostmodification such as *the fact that I love you; John Hunter, a salesman from Leeds;* . Often via appositions or relative clauses.
- Generic noun phrases: *I like **lions***;

- ***The badger*** *lives underground.* (generic starting with "the" if not referring to a specific badger).
- General mentions (not marked as old) even if they are repeated. ***Crime*** *should be tackled early. Otherwise,* ***crime*** *will take over.* (Here, *crime* is easily understandable on its own by lexical knowledge only).
- Most elliptical expressions, often marked by *one/ones* or similar expressions. Alternatively, they can just mean that the real noun is left out altogether from the expression. Examples are:
  - *John can play five instruments, and Mary can play* ***six***. (Here *six* is just new, not comparative (no explicit lexical marker such as other, such etc.). It is also not bridging as you cannot do a bridging rephrasal such as *six of the five instruments*.
  - *I need a new shirt. I want to buy* ***a white one***.

7 More examples, hints and precedence rules

7.1 The word *one/ones*

One can be used in many different meanings and one has to be careful. Some examples:

1. *Katja went to the beach* ***one fine day***. (Here, *one fine day* is a new entity)
2. Set membership bridging: *I studied three modules.* ***One module*** *I did not pass.*
3. Elliptical expressions: *I need a new shirt. I went for* ***a white one*** Here this is new.
4. Coreference: *I really liked Pulp Fiction. Normally I am not a fan of Quentin Tarantino but* ***that one*** *was great.* (should be premarked in the tool).

7.2 X itself: no mention

Most reflexive pronouns (*itself, themselves, himself* etc.) are premarked as `old` as they refer back to a previous entity such as in *He washed* ***himself***. However, in rare cases they are only used as an emphasiser in the construction *x itself/herself/...*. An example in context is *George Bush* ***himself*** *did believe that Iraq had WMD*. Here, *himself* is just an emphasiser and should be marked as a non-mention. This is the only grammatical construction with reflexives to be considered a non-mention.

7.3 Headlines

Headlines are sometimes deleted from the newspaper article in the corpus given. It is however still possible that the first sentence of the newspaper article contains a reference to that deleted headline such as

1. ***This company*** *announced a new . . .*

These cases are to be annotated as *new* as the headline has not been given.

7.4 Bylines

The last sentence of the text is often a so-called byline where the author is described. Sometimes in that case, the journal (Wall Street Journal) is mentioned with *the journal* or *the WSJ* or similar. In these cases, please annotated this as `mediated/world_knowledge` as the readers of this text would know what the journal is.

7.5 Time

Specific *and unique* times are either old or mediated world knowledge as they behave like named entities or "the" noun phrases. Examples are:

1. references to specific dates or years: *Tuesday; 16th of November; 1960* interpretable on their own.
2. references to specific dates or years interpretable uniquely by reference to the current time: *tomorrow; the next three weeks*

In contrast, time expressions that are not reference to a specific time or time phase, are normally `new`.

1. References to unspecified time periods: *this takes* ***three weeks***
2. Reference to generic times: *all the time, every day, every Tuesday, Tuesdays*

Of course, times can also be `comparative` (*it takes another three weeks*) or bridging in rare cases.

The different annotation of times effects annotation of `mediated/syntactic`.

1. *I went on* ***a three week holiday*** is new, as *three week* as a length of time is new.
2. *The August holiday* is mediated/syntactic if *August* refers to a specific August.

7.6 Money

Most money items are not unique references to specific money but just general sums, so are `new` as in

1. *I paid **2000 dollars**.*

However, they can also be `old`, `comparative`, `func` or `bridging` (among others) as in the examples below.

1. *I was paid 2000 dollars. I spent **the 2000 dollars** on a new car.* (`old`)
2. *He paid me **another 2000 dollars.*** (`comparative`)
3. *The IBM shares fell **3 cents** to **24 cents*** (both are `func`)
4. *I was paid 2000 dollars. I spent **1000 dollars** on a new car and saved the rest.* (`bridging/set`)

Again this affects `mediated/syntactic`.

7.7 Percentages

Percentages can be almost any category and the same rules as for money terms apply with some examples given below.

1. *He paid me only **80% of my deposit** back.* (`mediated/syntactic` as *my deposit* is mediated or old).
2. *I paid him a large deposit. He paid me only **80%** back* (`mediated/bridging`)
3. *My agent takes a 10% cut of my fees. Now he wants **another 5%*** (`mediated/comparative`)
4. *He likes **80% of sitcoms*** (new)

7.8 Other numbers (without head noun)

Other numbers without a non-number head noun without a clear case of bridging or syntactic mediation will be new.

1. *The game ended **3-2**.* (Here this is new)
2. *A **40 GB** disk* (again new)
3. *a **19-1** vote.* (Again new)

4. *I can speak three foreign languages.* ***Two*** *I learned at school.* (`bridging`: two of those three foreign languages)

5. *I can speek three foreign languages. I would rather speak* ***five****.* (`new` as it is an ellipsis, not *five of the three foreign languages*)

8 Check List for Frequently Made Errors

All of the check list items below are also mentioned in the main text. The descriptions in the main text include more details as well as examples than the check list.

8.1 Other_Coref

- ✓*the/this/that/these/those* noun phrases or pronouns referring to whole sentences or clauses.
- ✓Clear coreference between a singular and a plural bracketed noun phrase.
- ✓Clear coreference between a bracketed noun phrase and an aggregate of two or more previous noun phrases.
- ✓Forgotten pronoun annotation or repetition of proper names (only).
- ↯Coreference between two bracketed noun phrases with the same number (both singular or both plural) is not allowed.
- ↯Coreference between a bracketed noun phrase and a simple verb (phrase) previously is not allowed.

8.2 Comparative

- ✓Lexically marked
- ✓Antecedent can be inside or outside of text
- ✓Antecedent inside text can be noun phrase or verb, sometimes longer stretches of text like a whole clause/sentence.
- ✓Can in rare cases need an additional bridging link

8.3 Syntactic

- ✓Needs to have a bracketed entity inside that mediates
- ✓Mediation via PP, proper name premodifier, of-genitive, saxon genitive, possessive pronouns only.
- ↯Appositions (*John, my new boyfriend*) do not mediate.
- ↯Relative Clauses do not mediate
- ↯verbal complements (*the idea of going to London*) do not mediate
- ↯Adjective premodifiers do not mediate (*the American government*) — entity must be bracketed!
- ↯common noun premodifiers in compounds do not mediate (*the cave walls*)
- ✓Do not forget additional links if other mandatory links are missing.

8.4 Bridging

- ✓Mandatory role is missing
- ✓Antecedent can participate in a rephrasal sounding fine in the whole sentence.
- ✓Semantic type of antecedent and bridging entity are not the same (exception: set Bridging)
- ✓Several bridging links are possible
- ✓If antecedent has several instantiations within the text link to the one closest to the bridging entity.

8.5 World Knowledge or Text

- ✓For proper names and *the* NPs only
- ✓Unique within the whole world or the discourse context (but no coreference or bridging possible)