Anaphoric relations in the Copenhagen Dependency Treebanks

Iørn Korzen and Matthias Buch-Kromann*[†]
*Copenhagen Business School and [†]Copenhagen Business School

Abstract

The Copenhagen Dependency Treebanks (CDT) are a set of parallel treebanks for Danish, English, German, Italian, and Spanish. One of the main objectives of the CDT is to arrive at a unified description and annotation system for syntax, morphology, discourse, and anaphora. The treebanks are currently in the process of being annotated for these levels in all five languages. After a brief discussion of the subdivisions of the so-called bridging anaphors proposed by different scholars, we describe the classification and terminology adopted in the CDT. The main distinction here is the very common one between coreferential and associative anaphors, special attention being given to the latter group. Resumptive and evolving anaphors are treated as special subgroups of the coreferential anaphors. A list of the associative relations proposed by the CDT with authentic examples concludes the paper.

1 Introduction. The Copenhagen Dependency Treebanks

The purpose of this paper is partly to discuss the classification system and terminology adopted for anaphora by various scholars, and partly to describe the way anaphora is treated in the Copenhagen Dependency Treebanks. Special attention will be given to the "associative anaphors", which appear to be the most complex of the main anaphor types.

The Copenhagen Dependency Treebanks, CDT, are a set of parallel treebanks for Danish, English, German, Italian, and Spanish which are currently being annotated for syntax, morphology, discourse, and anaphora in all five languages.[1] The corpus consists of 100,000 words compiled from 200-250 word excerpts from Danish mixed-genre texts, which have been translated into the other languages by native translators. All 100,000 words have been translated into English, while 70,000 words have been translated into each of the other languages. All texts have been automatically annotated for parts of speech. A main objective of the CDT is to arrive at a unified description and annotation system for syntax, morphology, and discourse which at the same time can take cross-linguistic differences into account (Buch-Kromann *et al.* 2009).

After a brief terminological discussion in section 2, section 3 describes the distinction between the main anaphor types adopted in the CDT, and section 4 presents the CDT analysis of coreference. Sections 5 and 6 are dedicated to associative anaphora and section 7 to a few technical remarks.

2 "Bridging", "coreferential" and "associative" anaphors

The terms "bridge" and "bridging" (in the sense relevant to this paper) probably first appear in Clark (1975). Here, bridging is defined as the construction of the implicatures with which the listener bridges "the gap from what he knows to the intended

1 At this point, the anaphora analysis and annotation are confined to nominal anaphora.

Antecedent" (*op. cit.* 170). Clark includes "direct reference" (possibly with same-head NPs), "indirect reference by association", "indirect reference by characterization" (i.e. semantic roles), and the rhetorical relations "reasons", "causes", "consequences", and "concurrences" as situations that require an implicature "of some sort".

Subsequently, the term "bridging" has appeared frequently in the linguistic and computational literature, with more or less the same subclasses, except that coreferential pronouns and same-head NPs are generally left out, see e.g. Poesio *et al.* (1997, 2), Vieira and Poesio (2000, 558), and Caselli (2009, 73). In Vieira and Poesio (2000, 542), the "bridging descriptions" are summed up to be the "definite descriptions that either

(i) have an antecedent denoting the same discourse entity, but using a different head noun (as in *house . . . building),* or

(ii) are related by a relation other than identity to an entity already introduced in the discourse".

The same distinction, but expressed with the terms "coreferential" and "associative anaphors" respectively, is found in the work of a number of scholars, especially in the Romance tradition. Poesio and Vieira (1998, 187) cite Hawkins' (1978, 107/123) distinction between "Anaphoric Uses" and "Associative Anaphoric Uses", but in French the term "associative" is actually a lot older. It was probably first used as early as 1919 by Guillaume (1919, 162-163) but is now generally found in the theoretical linguistic literature, see e.g. Kleiber (1997a/b, and 2001), Schnedecker *et al.* (1994), Cornish (1999), Lundquist (2000), Korzen (2003 and 2009), and many others.[2]

In the last decade, also a number of schemes for anaphoric annotation have been released. Some of them confine themselves to coreference relations, e.g. the VENEX corpus (Poesio *et al.* 2004), the Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos 2007), and the Portuguese and French corpus analysed by Vieira *et al.* (2002). On the other hand, the analyses e.g. of the GNOME Corpus (Poesio 2004), the ARRAU Corpus (Poesio and Artstein 2008), the Dutch COREA corpus (Hendrickx *et al.* 2008), and the Italian Live Memories Corpus (Rodríguez *et al.* 2010) consider coreference as well as certain associative relations such as set membership, subset, ownership, and part-of relations. The Prague Dependency Treebank, PDT (Nedoluzhko *et al.* 2009) performs a wider range of bridging annotation including relations such as contrast, location–resident, relatives, and event–argument. Navarretta (2010) focuses on abstract pronominal anaphora in the DAD parallel corpora. Unlike most of the cited studies, which use automatic or semi-automatic annotation for instance of "markables", i.e. text constituents (mainly NPs and pronouns or pronominal phrases) that may enter in anaphoric relations,[3] all anaphor annotation is done manually in the CDT.

3 Main anaphor types in the CDT

In view of the very obvious differences between coreferential and associative relations,

2 For other designations of the "associative" anaphors in the theoretical linguistic literature, see e.g. Korzen (1996, 548-549).

3 The PDT has explicitly chosen not to use markables.

the same overall distinction has been retained in the CDT, whose aim it is to handle all nominal anaphor types in the two groups, thus, among the coreferential types, both same-head and non-same-head NP anaphors.

Graphically, the relation between text constituents and discourse referents (DRs)[4] in the two cases may be described as in Figure 1, where the dashed arrows indicate the "bridging", or relation deduction, undertaken by the hearer/reader, and the dotted double arrow in the case of the associative anaphor (part B of the Figure) indicates the "association" between the two discourse referents in question:

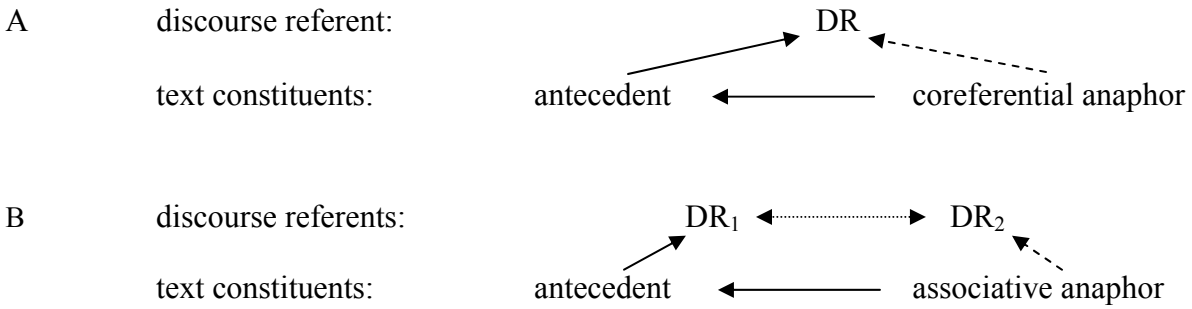

Figure 1. The relation between text constituents and discourse referents in the case of coreferential and associative anaphors.[5]

The so-called "evolving" anaphors refer to the same discourse referent as the antecedent, but after it has undergone radical changes in its ontological status, e.g.:

(1) The compactor crushed *a VW*. A huge crane then moved ***it*** to a railcar. (cit.: Asher 2000, 142).[6]

Therefore they can be seen as a sort of interface between coreference and associative anaphors, since the discourse referent is technically the same but markedly different.

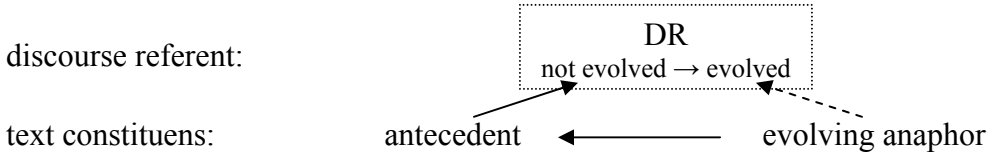

Figure 2. The relation text – extralinguistic context in the case of evolving anaphors.

In order to restrict the number of main anaphor types, the CDT treat evolving anaphors, as well as resumptive anaphors (which anaphorise whole sentences, clause or predicates, see Table 1 below and footnote 10), as special coreferential *sub*groups.

4 Discourse referents in the sense proposed by Karttunen (1969) and since then widely adopted in the literature.

5 In Webber's (1988) terminology, the coreferential anaphor specifies the referent which has already been evoked and specified by the antecedent, whereas the associative anaphor specifies **and** evokes its referent.

6 On "evolving anaphors" see also for instance Charolles and Schnedecker (1993), Korzen (2006), and Lundquist (2007).

4 Coreference in the CDT

In the CDT annotation, coreferential anaphors are subdivided partly according to their linguistic material (pronouns, same-head NPs and non-same-head NPs) and partly according to their semantic content (resumptive and evolving anaphors). The CDT annotation arrows go from antecedent to anaphor and – in the case of a longer anaphoric chain – from the last occurring anaphor to the new one:[7]

Coreferential anaphor (and cataphor) labels	**Examples (antecedent → anaphor)**
A. COREF: coreferential pronouns and other pro-forms	*a car → it/this; John Smith → he; you → you;*[8]
B. COREF-IDEN: coreferential NPs with lexical identity	*a car → the car; a big car → the/this big car;*
C. COREF-VAR: coreferential NPs with lexical variety	*a car → the vehicle; a yellow car → the/this car;*[9]
D. COREF-RES: resumptive anaphors	[a sentence, clause or predicate] → *the episode, this incident*[10]
E. COREF-RES.PRG: resumptive anaphors referring to a speech act	*"I shall be back tomorrow" → the threat, the promise, the statement*
F. COREF-EVOL: evolving anaphors	see example (1) above

Table 1. Coreferential anaphor (and cataphor) types.

See also Figure 5 in section 7. The distinction between A, B and C concerns the linguistic material, whereas D, E and F are special semantic subgroups. D and E will also be either COREF (in the case of a pronoun) or COREF-VAR (in the case of an NP), and F will be either COREF, COREF-IDEN or COREF-VAR,[11] but this is not specified in the annotation. Cases of repeated proper nouns, e.g.:

(2) a. *John Smith → John Smith*
b. *John Smith → John*

are included as cases of COREF-IDEN (2a) or COREF-VAR (2b), even if they differ from common noun anaphors by not being necessarily dependent on their antecedent. Such cases, as well as repeated deictic pronouns, can easily be found and studied in searches that combine the anaphor label and the part-of-speech label.[12]

7 The subgroups of Table 1 include both anaphors and cataphors. In the case of cataphors, the arrows go from postcedent to cataphor.

8 Similarly, we annotate anaphoric relations between the subject of a verb of saying and a coreferential pronoun in the direct or indirect speech, and between coreferential pronouns in the different parts of dialogue, e.g. A: *...I...* → B: *...you...*

9 If, in a longer anaphoric chain, there is a relation between a pronoun, e.g. *he* (as last occurring anaphor), to an NP, e.g. *John Smith* (as new anaphor), this relation will be labelled COREF-VAR.

10 Typical resumptive anaphors are e.g. nominalizations, gerunds, and scene descriptions, and they can be subdivided in neutral NPs (e.g. *the operation, the activity, the situation*), NPs that either interpret or evaluate the story line (e.g. *the damage, the misdeed, the error*), or NPs that refer to the plot or story structure (e.g. *the scene, the gag, the comedy*); for more detail see also Korzen (2007).

11 On the material of the evolving anaphors, see especially the references mentioned in footnote 6. For more detail on the CDT annotation system, including the CDT manual, see the references in footnote 22.

12 Regarding the CDT search possibilities with the aid of the DTAG annotation tool, see Buch-Kromann *et al.* (2009).

Our COREF-VAR group is very heterogeneous, at this point in time containing both cases of different (common or proper) nouns and different attributives. In due course, we may decide to subdivide this group into more homogeneous subgroups.

5 Association and the Generative Lexicon

As is well-known, the ways in which two discourse referents may "associate", as illustrated by the dotted double arrow in Figure 1B above, have been discussed extensively in the literature in the last few decades (as well as by Guillaume 1919, 162ff.). Especially after the appearance of Pustejovsky's (1995) "Generative Lexicon", a number of scholars have seen the prospects of uniting lexical generativity, or "entailments" (Bos *et al.* 1995, 2), with the phenomenon of associative anaphors; see e.g. Bos *et al.* (*op. cit.*); Lundquist (2000); Henry & Bassac (2008); Caselli (2009); Korzen (2000; 2003; 2009). Particularly useful is Pustejovsky's "qualia structure" with the four qualia, or roles, attributable to any artefact[13]:

A. FORMAL: That which distinguishes the object within a larger domain (orientation, magnitude, shape, dimensionality, color, position).
B. CONSTITUTIVE: The relation between an object and its constituents, or proper parts (material, weight, parts and component elements).
C. AGENTIVE: Factors involved in the origin or "bringing about" of the object.
D. TELIC: Purpose and function of the object.

Figure 3. Pustejovsky's (1995, 76ff / 85ff) "Qualia Structure".

Each of the four roles contains either entities/elements (A-B) or events (C-D) potentially generatable in a "default form", and both such entities/elements and events on the one hand and the arguments of the events on the other may be activated in an association relation to the object in question, i.e. function as associative anaphors to the NP designating this object.

But also before Pustejovsky, there were similar attempts to combine an apparent lexical and cognitive associability between concepts. For instance Hawkins (1978, 123-124) mentions "part-of relationship" and "attributes of an object" as possible "triggers" of associative anaphoric relations, such as for instance – with reference to *a car – the wheels, the steering-wheel, the passenger seats* and *the length, the colour, the weight,* corresponding to Pustejovsky's constitutive and formal qualia respectively.

Löbner (1998) distinguishes between sortal, relational and functional concepts, the functional ones being those that denote a 1-to-1 relation to a referent. He adds (*op. cit.* 4) that "all sortal nouns also encode relational or functional characteristics". Even a prototypical sortal noun like *book* has "a meaning that relates its possible referents to ways in which one can interact with books: write them, read them, […] etc.", a meaning

13 Natural objects have a FORMAL and a CONSTITUTIVE quale, but not an AGENTIVE or a TELIC quale.

that corresponds to Pustejovsky's agentive and telic qualia. Functional Concepts which have a possessor argument[14] are claimed to underlie all definite associative anaphors, whereas relational nouns which do not denote a 1-to-1 relation, e.g. *finger, hand, son, aunt*, are rejected as possible associative anaphors (*op. cit.* 10-11). This, however, is not necessarily true, as the following (very typical) Italian and Danish examples will show:

(3) Disse tutto questo e altro, che non ricordo. Mentre parlava, neppure io lo guardavo. [...] D'un tratto *mi* posò ***la mano*** su***l braccio***. "Avrei bisogno che tu mi dessi un consiglio", fece. (Giorgio Bassani, *Gli occhiali d'oro.* Oscar Mondadori, Verona, 1973, p. 139)
'He said this and other things that I don't recall. As he spoke, I didn't even look at him. […] [lit.:] Suddenly *me* he put ***the hand*** on ***the arm***. "I need you to give me some advice", he said.'

(4) Politiet affyrede to skud mod *manden*. Det ene ramte *ham* i ***låret***. (Danish TV2-news 17.4.99)
'The police fired two shots at *the man*. One hit *him* in ***the thigh***.'

Even if *la mano* 'the hand'*, il braccio* 'the arm' and *låret* 'the thigh' are all in the singular, there is no reason to believe that the people involved are mutilated, and certainly an expression such as *hit him in* followed by a singular form of a noun denoting body parts we (normally) have more than one of is very common in English as well, as a few searches on Google reveal.

Associative anaphors tend to appear particularly often in the Romance languages, where for instance a case such as (5) is quite common:

(5) In questo momento *Fiorenza* non c'è. Io sono ***la figlia***.
[lit.:] 'At the moment *Fiorenza* is not here. I am ***the daughter***.'

Example (5) is based on an authentic example cited and discussed in Korzen (1996, 518, see also p. 35-36), and in "real life" the person "Fiorenza" actually has two daughters.

Kleiber (1997a/b; 2001, 263-367) operates with the following typology of four main groups of associative anaphors:

A. MERONYMIC: the anaphor is a fixed part of the antecedent, e.g. *a car → the wheel, a cup → the handle.*
B. LOCATIVE: the anaphor is located in the antecedent, e.g. *a village → the church, a kitchen → the refrigerator.*
C. FUNCTIONAL: the anaphor fulfils a function in relation to the antecedent, e.g. *a town → the mayor, a restaurant → the waiter.*
D. ACTANTS: the anaphor has a semantic and/or syntactic role in relation to a predicative antecedent, e.g. *an operation → the surgeon / the patient* (arguments)*, he cut the bread and put away* ***the knife*** (instrument).

Figure 4. Kleiber's (1997a/b; 2001, 263-367) typology of associative anaphors.

14 Such concepts are said always to have a situational argument as well and are therefore termed FC2s (*op. cit.* 5).

Of these, the A, C and some of the D types are covered by Pustejovsky's qualia. One could argue that some of Kleiber's types are overlapping: a meronymic anaphor is also located in the antecedent and may very well fulfil a function as well. We shall return to these (thorny) problems below.

6 The associative anaphors in the CDT

The CDT classification and subdivision of associative anaphors are highly inspired by the typologies mentioned in the previous section, as the following lists will show. Since CDT is an ongoing project in which we by and large have worked, and are working, empirically, we cannot exclude that further analyses will give rise to changes, but we believe they will be minor. In the following text examples (all authentic), the antecedents are printed in italics and the anaphors in bold italics followed by the label. A number between parentheses following the example indicates the number of the text in the CDT corpus. In a few cases, text examples come from other sources. Unlike the coreferential anaphors, the structure of the associative labels is hierarchic, which means that the following types are all associative subtypes. As in the CDT syntax and discourse annotation (and inspired by the Penn Discourse Treebank), this means that in case of uncertainty, the annotator can remain on a higher (more generic) annotation level.[15]

With a few exceptions (see footnote 16), associative anaphors seem classifiable according to two parameters:

- lexical semantics and generativity, qualia structure;
- semantic roles in relation to a predicate; the predicate may be either directly expressed by the antecedent or generatable from it.

1. Qualia structure	2. Semantic roles	3. Other types[16]
ASSOC-FORMAL	ASSOC-AGENT	ASSOC-LOC(ation)
ASSOC-CONST(itutive)	ASSOC-PATIENT	ASSOC-TIME
ASSOC-AGENTIVE	ASSOC-EXPER(iencer)	ASSOC-EVENT
ASSOC-TELIC	ASSOC-REC(ipient)	
	ASSOC-INST(rument)	

Table 2. Associative subtypes, parameters and labels.

6.1 The anaphor is associated with the antecedent with regard to its qualia structure

ASSOC-FORMAL

The FORMAL quale expresses static information about the object's characteristics. If the

15 A similar solution does not seem to be needed in the case of the coreferential anaphors, where our subdivision should not give rise to much uncertainty. At the most, a COREF-RES or a COREF-EVOL anaphor might risk a categorization as a COREF, a COREF-IDEN or a COREF-VAR anaphor, which would neither be catastrophic, nor untrue.

16 In fact, these "exceptions" may be seen as extensions of the other two subtypes. However, LOCATION and TIME are labelled as semantic roles by some scholars, see e.g. Larson (1984, 202), and EVENT expresses a predication linked to the antecedent, similar to but more generic than the TELIC and AGENTIVE qualia. See 6.3 below.

anaphor is associated with the antecedent with regard to its FORMAL quale, it may designate the shape, dimension, colour, etc. of the object designated by the antecedent:

(6) The ham to be used in the dish must not be too salty. You cannot use *the thin slices*, which are packaged in the refrigerated counter. *They* are too salty and too wet and ***the flavour*** [ASSOC-FORMAL] is not good enough. (148)

The other three qualia roles contain information about relations that the object referred to by the antecedent can be a part of, i.e. they constitute predicates of which the antecedent is an argument. In these cases, an associative anaphor can function as the other argument or as the predicate itself.

ASSOC-CONST

Also the CONSTITUTIVE quale expresses static information about the object (parts, elements, material, content, etc.). The predicates of which antecedent and anaphor are arguments are *has_part, consists_of, is_part_of,* and the like. In (7) the anaphor is part of the antecedent, in (8) vice versa. In both cases we talk about ASSOC-CONST-relations:[17]

(7) The accident took place at dinner time around 6:45 p.m. last night […]. I saw *the plane* with its nose pointing downward, ***the left wing*** [ASSOC-CONST] up and ***the right wing*** [ASSOC-CONST] down over behind the flat building. (1536)

(8) On September 8, DE BEERS CENTENARY opened an office in *Moscow*. Present were also De Beers' top people, Russian politicians, diplomats and representatives of ***the country***'s [ASSOC-CONST] diamond industry and trade. (431)[18]

ASSOC-TELIC and ASSOC-AGENTIVE

If the anaphor is associated with the antecedent with regard to its AGENTIVE or TELIC quale, the anaphor may designate the quale predicate itself or an inferable argument of such a predicate. Examples (9) and (10) are cases of predicative anaphors:

(9) As previously explained, we were waiting for an approval from Sony as we submitted to them *a new version of Blood Bowl PSP*. […] *This new version* has been finally approved and ***the production*** [ASSOC-AGENTIVE] started. Please find below the list of fixes that were made. (http://www.gamefaqs.com/boards/944028-blood-bowl/52159350, accessed October 8th, 2010)

(10) However, not all debriefings are held after the simulation, but in *certain instances*, for example, where ***the aim*** [ASSOC-TELIC] is to teach a technical skill […] debriefing may occur during the simulation, in-scenario debriefing.

17 "The constitutive […] quale refers not only to the parts or material of an object, but defines, for an object, what that object is logically part of, if such a relation exists. The relation PART-OF allows for both abstractions" (Pustejovsky 1995, 98).

18 It may be debatable whether the antecedent is *Russian* rather than *Moscow,* but they can both function as antecedents, which can be proved in a simple test where one or the other is omitted from the co-text. We should also add that in cases like this, a precise borderline between ASSOC-CONST and ASSOC-LOC can be very hard to draw; see below.

(citeseerx.ist.psu.edu/viewdoc/download, accessed October 13th, 2010)

Anaphors that designate a particular semantic role of the given quale predicate are treated as subtypes. The precise analysis of the role in question will depend on the inferred predicate. Thus, in these cases the annotators are asked to add the inferred predicate between parentheses. As regards AGENTIVE subtypes, so far we have only encountered the semantic role AGENT:

(11) In April 2003, marking the tenth anniversary of the Waco Massacre, *a new film* was released. According to ***the producer*** [ASSOC-AGENTIVE.AGENT/(produce)], "Waco: A New Revelation" is a film so disturbing that […] it triggered new investigations in both houses of Congress […]. (http://www.serendipity.li/waco.html, accessed September 5th, 2010)

In (11), in order to infer *the producer*, we must first activate the agentive quale *produce.* Similarly, in (12) and (13) *the pilot* and *both apprentices* can be seen as the semantic roles AGENT and PATIENT of the telic qualia of a *flight* (i.e. *to fly*) and a *test* (i.e. *to examine*) respectively:

(12) The accident took place at dinner time around 6:45 p.m. last night, shortly after *the El-Al flight* […] lifted off from Amsterdam's Schiphol airport.
The pilot [ASSOC-TELIC.AGENT/(fly)] suddenly reported to the control tower that he had engine problems […]. (1536)

(13) *Two journeyman tests* were passed in August. ***Both apprentices*** [ASSOC-TELIC.PATIENT/(examine)] are trained at the Royal Copenhagen A/S Georg Jensen Silversmithy. (431)

In some cases, more than one subtype interpretation may apply, for which reason an annotator could remain at the ASSOC-TELIC level[19]:

(14) The men in question are simply film reviewers and quite harmless. […] If some nonsense should sometimes appear in a *film review*, it is thus due not to time pressure, even though, of course, it is most convenient for the reviewers if ***the readers*** [ASSOC-TELIC.AGENT/(read) or ASSOC-TELIC.REC/(receive)] believe that. (647)

6.2 The antecedent is predicative and the anaphor is a semantic role

If the antecedent is a predicate or a predicative noun, the anaphor can constitute a semantic role which is related to it directly, not (necessarily) via a quale:

(15) *The operation* itself requires general anesthesia ... the patient is asleep for the entire course of the operation. ***The surgeon*** [ASSOC-AGENT] opens the chest by dividing the breast bone or sternum. (http://www.heartsurgeons.com/pr3.html, accessed August 5th, 2010)[20]

19 With the risk of confusion with cases such as (10). At this point, we have not been able to solve this problem.
20 The tree dots appeared as shown in the cited text.

(16) *The operation* itself requires general anesthesia ... ***the patient*** [ASSOC-PATIENT] is asleep for the entire course of the operation. The surgeon opens the chest by dividing the breast bone or sternum. (http://www.heartsurgeons.com/pr3.html, accessed August 5th, 2010)

(17) *The accident* took place at dinner time around 6:45 p.m. last night […]. "[…] The pilot attempted to right the plane - then I could not see more, but suddenly there were sparks in the air," says ***eyewitness Peter de Neef*** [ASSOC-EXPER]. (1536)

(18) "[…] This is *the most violent attack* to this point. ***The bombs*** [ASSOC-INST] fell half a mile from the hotel," reported John Hollimann […] (61).

6.3 Other types

According to the definition of "semantic roles" (see footnote 16), TIME and LOCATION may belong to the previous section or they may be extensions of it. An ASSOC-TIME anaphor may indicate a point in time linked to the antecedent, which may be a predicate or predicative noun, another time indication, as in (19), or a more general narrative frame, as in (20):

(19) As mentioned, the season will begin on *March 16* with the showdown between AGF and Brøndby, followed ***the day after*** [ASSOC-TIME] by games between: Ikast-Lyngby, B 1903-Silkeborg, AaB-Vejle and FremOB. (43)

(20) Aspiring chef dies hours after making ultra-hot sauce for chilli-eating contest [headline]
Andrew Lee made an ultra-hot sauce with homegrown chillis. ***The morning after*** [ASSOC-TIME] he was found unconscious and paramedics were unable to revive him.
(Mailonline, http://www.dailymail.co.uk/news/, accessed August 6th, 2010)

The ASSOC-LOC relation is very close to the ASSOC-CONST relation, and a precise borderline can be hard to draw. An ASSOC-LOC anaphor is located in the antecedent (or vice versa) without being necessarily a constitutive part:

(21) Upon entry, the officers saw *the kitchen* with many dirty dishes, spoiled food on the floor and in ***the refrigerator***, and bags of trash and other combustibles on top of ***the stove***. (http://www.leagle.com/xmlResult.aspx?xmldoc=197621858CalApp3d160_1205.xml&docbase=CSLWAR1-1950-1985, accessed November 10th 2010).

Similarly, as an extension of examples (9) and (10), a predicative anaphor may express an event which is associable with the antecedent, but not necessarily with regard to its qualia structure. In such cases we adopt the more generic label ASSOC-EVENT:

(22) Hamid Jafar was very eager to show his appreciation of the agreement to his *Iraqi* partners. Shortly before ***the invasion*** [ASSOC-EVENT], he ordered an engraved, Swiss, gold pistol assessed at 7,000 pounds from […] the English Queen's jeweller in London. (939)

7 Graphs and inter-annotator agreement

The CDT graphs are generated with the DTAG annotation tool described in Kromann (2003)[21] and use directed edges with the relation labels shown at the arrow head. Figure 5 shows the syntax annotation (above the nodes) and anaphor annotation (below the nodes) of the last sentence of example (7).

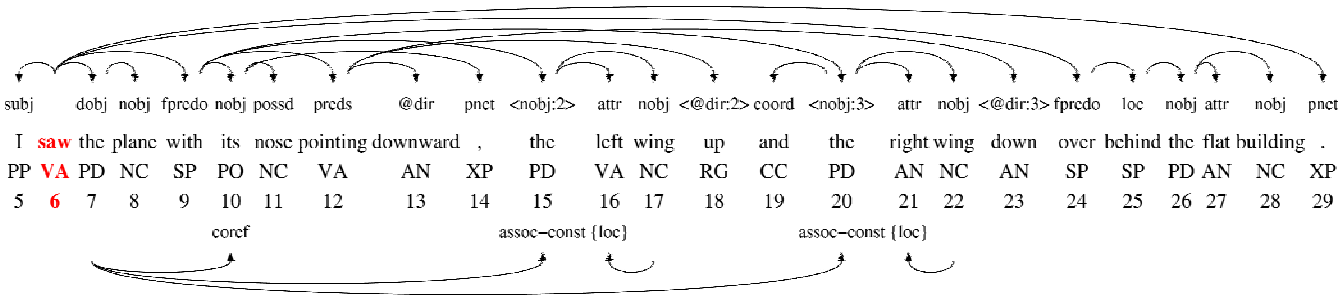

Figure 5. A CDT syntax and anaphor annotation of the sentence *I saw the plane with its nose pointing downward, the left wing up and the right wing down over behind the flat building.*

The annotation shows that the NP *the plane* (nodes 7-8)[22] is the antecedent of a coreferential pronoun (node 10) and two ASSOC-CONST anaphors (nodes 15-17 and 20-22). In this figure, we have omitted most of the secondary semantic relations, also annotated below the text, but left two of them behind (nodes 16 and 21) in order to give an impression of this annotation category. Even if they are annotated below the text like the anaphoric relations, secondary semantic relations clearly belong to a different linguistic dimension, just like the syntactic and discourse relations belong to different dimensions although they are both annotated above the text in the CDT.

In order to test our anaphor relation system by computing inter-annotator agreement as soon as possible, 25 texts have been annotated independently by two annotators. The texts contained a total of 466 anaphoric relations, and Table 3 (taken from the CDT manual, Buch-Kromann *et al.* 2010) shows the level of inter-annotator agreement and the frequency of the anaphoric relations found in the 25 texts. Agreement is reported as percentage agreement[23] in the following way:

- *Full labelled agreement, A:* the probability that another annotator assigns the same label and out-node to the relation;
- *Unlabelled agreement, A_U:* the probability that another annotator assigns the same out-node (but not necessarily the same label) to the relation;

21 More references can be found at http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT.

22 Of which the determiner is considered head and the lexical noun nominal object, "nobj". For more detail on CDT graphs, analyses, and annotation, see Buch-Kromann *et al.* (2009, 2010). The CDT-manual can be downloaded from the URL of the latter reference: http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf.

23 The estimated level of agreement is defined as the probability that another annotator assigns the same label and/or out-node to the relation (this number may be inaccurate if the relation count is small). We do not report chance-corrected scores because they are harder to interpret and their usefulness is contested (Reidsma and Carletta, 2008; Buch-Kromann, 2010). For more detail, we refer our readers to the CDT manual.

- *Label agreement, A_L:* the probability that another annotator assigns the same label (but not necessarily the same out-node) to the relation.

Relation name	Agreement % $A - A_U - A_L$	Relation count	Relation name	Agreement % $A - A_U - A_L$	Relation count
COREF	84 – 85 –92	141	ASSOC (subtype)	39 – 83 – 39	9
COREF-VAR	71 – 79 –79	97	ASSOC-LOC	100 – 100 – 100	5
REF[24]	100 – 100 – 100	63	ASSOC-AGENTIVE	25 – 50 – 50	4
COREF-IDEN	77 – 83 – 81	53	ASSOC-EVENT	100 – 100 – 100	3
ASSOC-CONST	59 – 77 – 67	39	ASSOC-FORMAL	100 – 100 – 100	1
COREF-RES	65 – 73 – 72	25	COREF-RES.PRG	0 – 0 – 0	1
ASSOC-TELIC	71 – 88 – 83	24	COREF-EVOL	0 – 100 – 0	1
			TOTAL	77 – 84 – 84	466

Table 3. Inter-annotator agreement based on 25 CDT texts with 466 anaphoric relations.

As a first test at a relatively early point in time, and considering that we include all nominal anaphors, even the most complex associative types, we find the result acceptable. However, we feel confident that an even better result can be obtained after more time for discussion and analysis together with the two annotators.

8 Conclusion

In this paper, we have described the anaphor annotation system in the Copenhagen Dependency Treebanks, an on-going project which is still in its relatively early stages. The over-all distinction is the very common one between coreference and associative anaphors, of which the latter group is clearly the most complex and complicated one. Associative anaphora has to do with how concepts relate to or associate with each other, and in this connection we have found it fruitful to look at lexical generativity and semantic association. A combination of Pustejovsky's qualia structure and the most common semantic roles (in Table 4 "semroles") played by arguments in connection with their predicates seems to be able to account for almost all cases of associative anaphora. The CDT project operates with a hierarchic label system that allows annotators to remain at a higher level in case of uncertainty as to subtypes. The ASSOC types and subtypes are the following:

24 REF regards syntactically determined coreference, typically used in relative clauses with a relative pronoun.

ASSOC-QUALIA (± semrole subtype):	ASSOC-SEMROLE:	ASSOC-OTHER:
ASSOC-FORMAL	ASSOC-AGENT	ASSOC-EVENT
ASSOC-CONST	ASSOC-EXPER	ASSOC-LOC
ASSOC-AGENTIVE	ASSOC-INST	ASSOC-TIME
ASSOC-AGENTIVE.AGENT	ASSOC-PATIENT	
ASSOC-TELIC	ASSOC-REC	
ASSOC-TELIC.AGENT		
ASSOC-TELIC.EXPER		
ASSOC-TELIC.INST		
ASSOC-TELIC.PATIENT		
ASSOC-TELIC.REC		

Table 4. Associative anaphora in the CDT.

At a later stage, cross-linguistic alignment will allow us to compare anaphoric relations in our five languages with great accuracy. For instance, it will enable us to identify and precisely describe the considerable typological differences between associative relations in Romance and Germanic languages, some of which were briefly illustrated in examples (3) and (5).

9 Acknowledgments

This work was supported by grants from the Danish Research Council for the Humanities and the Copenhagen Business School. We thank Lotte Jelsbech Knudsen and Morten Gylling-Jørgensen for many fruitful discussions and the anonymous reviewers for their useful comments.

References

Nicholas Asher. Events, Facts, Propositions, and Evolutive Anaphora. In James Higginbotham, Fabio Pianesi and Achille C. Varzi (eds). *Speaking of Events*. Oxford University Press, New York & Oxford, pages 123-150, 2000.

Johan Bos, Paul Buitelaar, and Anne-Marie Mineur. Bridging as Coercive Accommodation. In *Workshop on Computational Logic for Natural Language Processing (CLNLP)*, Edinburgh, 1995.

Matthias Buch-Kromann. Open challenges in treebanking: some thoughts based on the Copenhagen Dependency Treebanks. Invited paper at the Annotation and Exploitation of Parallel Corpora Workshop, Tartu, December 1-2, 2010.

Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. Uncovering the 'lost' structure of translations with parallel treebanks. In Inger M. Mees, Fabio Alves, and Susanne Göpferich, (eds), *Methodology, Technology and Innovation in Translation Process Research. Copenhagen Studies in Language* 38, Samfundslitteratur, Copenhagen, pages 199-224, 2009.

Matthias Buch-Kromann, Morten Gylling-Jørgensen, Lotte Jelsbech Knudsen, Iørn Korzen, and Henrik Høeg Müller. *The inventory of linguistic relations used in the Copenhagen Dependency Treebanks. Technical report. (The CDT manual).* Center for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, 2010. http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf

Tommaso Caselli. Using a Generative Lexicon Resource to Compute Bridging Anaphora in Italian. *Procesamiento del Lenguaje Natural* 42, pages 71-78, 2009.

Michel Charolles and Catherine Schnedecker. Coréférence et identité. Le problème des référent évolutifs. In *Langages* 112, pages 106-126, 1993.

Francis Cornish. *Anaphora, Discourse, and Understanding. Evidence from English and French.* Clarendon Press, Oxford, 1999.

Herbert H. Clark. Bridging. In R. C. Schank & B. L. Nash-Webber (eds), *Theoretical Issues in National Language Processing*. MIT, 1975.

Gustave Guillaume. *Le problème de l'Article e sa solution dans la Langue française*. Librairie Hachette, Paris, 1919. [Réédition Libraire A.-G. Nizet, Paris / Les Presses de l'Université Laval, Quebec, 1975].

John A. Hawkins. *Definiteness and Indefiniteness. A Study in Reference and Grammaticality Prediction.* Croom Helm, London, 1978.

Iris Hendrickx *et al.* A Coreference Corpus and Resolution System for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 144-149, 2008.

Patrick Henry and Christian Bassac. A toolkit for a Generative Lexicon. In *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris 2007, 2008.

Lauri Karttunen. Discourse Referents. In *International Conference on Computational Linguistics, COLING, Preprint No. 70*, 1969.

Georges Kleiber. Des anaphores associatives méronymiques aux anaphores associatives locatives. In *Verbum* XIX/1-2, pages 25-66, 1997a.

Georges Kleiber. Les anaphores associatives actantielles. In *Scolia* 10, pages 89-120, 1997b.

Georges Kleiber. *L'anaphore associative*. Presses Universitaires de France, Paris, 2001.

Iørn Korzen. *L'articolo italiano fra concetto ed entità. Vol. I-II*. [*Etudes Romanes* 36], Museum Tusculanum Press, Copenhagen, 1996.

Iørn Korzen. Pragmatica testuale e sintassi nominale. Gerarchie pragmatiche, determinazione nominale e relazioni anaforiche. In Korzen and Marello (eds), 2000, pages 81-109, 2000.

Iørn Korzen. Anafore e relazioni anaforiche. Un approccio pragmatico-cognitivo. In *Lingua nostra* LXII (3-4), pages 107-126, 2001.

Iørn Korzen. Anafora associativa: aspetti lessicali, testuali e contestuali. In Nicoletta Maraschio and Teresa Poggi Salani (eds). *Italia linguistica anno Mille, Italia linguistica anno Duemila.* Bulzoni, Roma, pages 593-607, 2003.

Iørn Korzen. Tipologia anaforica: il caso della cosiddetta "anafora evolutiva". In *Studi di grammatica italiana.* Accademia della Crusca, Firenze, XXV, pages 323-357, 2006.

Iørn Korzen. Linguistic typology, text structure and anaphors. In Korzen and Lundquist (eds), 2007, 93-109.

Iørn Korzen. Anafora associativa: ulteriori associazioni. In Federica Venier (ed.). *Tra pragmatica e linguistica testuale. Ricordando Maria-Elisabeth Conte.* [*Gli argomenti umani* 13]. Edizioni dell'Orso, Alessandria, pages 307-326, 2009.

Iørn Korzen and Carla Marello (eds). *Argomenti per una linguistica della traduzione / On linguistic aspects of translation / Notes pour una linguistique de la traduction. Gli argomenti umani* 4. Edizioni dell'Orso, Alessandria, 2000.

Iørn Korzen and Lita Lundquist (eds). *Comparing Anaphors. Between Sentences, Texts and Languages. Copenhagen Studies in Language,* 34. Samfundslitteratur Press, Copenhagen, 2007.

Olga Krasavina and Christian Chiarcos. PoCoS – Potsdam Coreference Scheme. In *LAW '07 Proceedings of the Linguistic Annotation Workshop*, 2007.

Matthias Trautner Kromann. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), 14-15*

November, Växjö, pages 217–220, 2003.
Mildred I. Larson. *Meaning-based translation. A guide to cross-language equivalence*. Lanham, New York / London, 1984.
Sebastian Löbner 1998. Definite Associative Anaphora. (manuscript) http://user.phil-fak.uni-duesseldorf.de/~loebner/publ/DAA-03.pdf
Lita Lundquist. Translating Associative Anaphors. A Linguistic and Psycholinguistic Study of Translation from Danish into French. In Korzen and Marello (eds) 2000, 111-129, 2000.
Lita Lundquist. Comparing evolving anaphors in Danish and French. In Korzen and Lundquist (eds), pages 111-125, 2007.
Costanza Navarretta. The DAD parallel corpora and their uses. In *Proceedings of LREC 2010, Malta, 17-23 May 2010*, pages 705-712, 2010.
Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*, pages 108–111, 2009.
Massimo Poesio. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*, 2004.
Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In Proceedings of the LREC 2008, Marrakech, Morocco, 2008.
Massimo Poesio and Renata Vieira. A corpus-based investigation of definite description use. In *Computational Linguistics* 24(2), pages 183-216, 1998.
Massimo Poesio, Renata Vieira and Simone Teufel. Resolving Bridging References in Unrestricted Text. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, Madrid, Spain, pages 1-6, 1997.
Massimo Poesio, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran, and Sara Tonelli. The VENEX corpus of anaphora and deixis in spoken and written Italian, 2004. http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf
James Pustejovsky. *The Generative Lexicon: A theory of computational lexical semantics.* MIT Press, Cambridge, MA, 1995.
Dennis Reidsma and Jean Carletta. Reliability measurement without limits. In *Computational Linguistics* 34(3), pages 319-326, 2008.
Kepa J. Rodríguez, Francesca Delogu,Yannick Versley, Egon W. Stemle, and Massimo Poesio. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, pages 157-163, 2010.
Catherine Schnedecker and Michel Charolles. Les référents évolutifs: points de vue ontologique et phénoménologique. In *Cahiers de linguistique française* 14, pages 197-227, 1993.
Catherine Schnedecker, Michel Charolles, Georges Kleiber, and Jean Davis (réd.). *L'anaphore associative. (Aspects linguistiques, psycholinguistiques et automatiques)*. Klincksieck, Paris, 1994.
Renata Vieira and Massimo Poesio. An Empirically-Based System for Processing Definite Descriptions. In *Computational Linguistics* 26(4), pages 539-593, 2000.
Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the DAARC, Estoril,* 2002.
Bonnie Webber. Tense as Discourse Anaphora. In *Computational Linguistics* 14 (2), pages 61-73, 1988.