AnCora-CO:
Coreference Guidelines
for Catalan and Spanish

Version 2.1

CLiC – CENTRE DE LLENGUATGE I COMPUTACIÓ

UNIVERSITAT DE BARCELONA

September 2008

CONTENTS

1. INTRODUCTION

Natural Language Processing (NLP) applications such as information extraction, text summarization and question answering need to identify all the information that is said about one same entity throughout a discourse in order to have a good text understanding. The following piece of news (1) illustrates a coreference chain with different linguistic expressions. [1]

> *La religiosa que es va curar gràcies a Joan XXIII defuig les aparicions públiques. Caterina Capitani es va refer d'una malaltia mortal. La beneficiària del miracle ha estat una de les estrelles de la cerimònia d'aquest matí a la plaça de Sant Pere. Però fins ara Ø s'ha mantingut amagada. [...] Una altra monja, la germana Adele, va començar a encomanar-la al papa bo.*

Following the terminology of the Automatic Content Extraction (ACE) program, we distinguish between *mentions* and *entities*:

- **Mention** is an instance of reference to an object.
- **Entity** is the collection of mentions referring to the same object in a document.

Thus, the above example contains five different mentions referring to the entity *the nun Caterina Capitani*. The entity is first evoked in the discourse by means of a definite description (*la religiosa que es va curar gràcies a Joan XXIII*) and it is next referred to in the following sentence by a proper noun (*Caterina Capitani*). Subsequently it is expressed via an anaphoric description (*la beneficiària del miracle*) –the previous proper noun helps complete its referential meaning. The last mentions take the form of a zero subject pronoun, and a clitic personal pronoun (*-la*).

We follow the discourse model by (Webber 1978), according to which coreference and anaphora occur between discourse entities (henceforth DE), which may or may not refer to specific objects in the real world. Two DEs are coreferential if they have the same discourse referent. It may be that one of them is anaphoric (2a), that is, interpreted with the aid of some previous DE (the antecedent), or it may be that both DEs are autonomous (2b). Thirdly, an entity can also be anaphoric if it is *colexical* with a previous item (2c): they share the semantic type, but they are not coreferential.

(2) a. Als treballadors ja no els queda res a reivindicar.
b. La capital de França...a París...
c. A algú se li acudirà organitzar festes de lluita de classes, igual que existeixen les de moros i cristians.

Coreference and anaphora are thus closely interrelated, although not all coreferential relations are anaphoric (2b), nor are all anaphoric relations coreferential (2c). We focus on annotating coreference links at both a specific and a generic[2] level, as well as two subsidiary semantic links on which we will comment in due time.

This manual presents the guidelines for the coreference annotation of the Spanish and Catalan AnCora corpora. Section 2 briefly introduces the corpus to be annotated and outlines the annotation process for coreference. Sections 3 and 4 are the core of the manual, specifying the attributes of mentions and the construction of entities (i.e. coreference chains), respectively. Section 5 includes the list of references.

[1] Following the syntactic annotation already contained in the corpus, zero pronouns are represented as **0**.

[2] As suggested by Carlson (1977) and Lyons (1999), we view generics either as proper names of classes, or proper names as a kind of generic including only one entity. Consequently, we contemplate coreference relations between generic NPs.

2. OVERVIEW

AnCora results from two different corpora: AnCora-Ca, the Catalan corpus, and AnCora-Es, the Spanish one. AnCora was built in an incremental way from the previous 3LB and CESS-ECE corpora, which come mostly from newspaper and newswire articles. AnCora-Ca consists of 75,000 words from the EFE news agency, 225,000 words from the ACN Catalan newswire agency and 200,000 words from the Catalan version of *El Periódico* newspaper. AnCora-Es contains 75,000 words from the Lexesp corpus, 225,000 words from the EFE Spanish newswire agency, and 200,000 from the Spanish version of *El Periódico* newspaper.

AnCora is the largest multilayer annotated corpus of Catalan and Spanish freely available from http://clic.ub.edu/ancora. The two corpora consist of half million words annotated (XML markup) at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, NEs, and WordNet nominal senses). AnCora is comparable to other corpora which are being developed at present: the English OntoNotes, and the Czech Prague Dependency Treebank.

The AnCora coreference annotation scheme has been inspired on the general criteria offered by the MATE meta-scheme (Poesio 2000; Poesio 2004). It involves two major tasks:

1) Identification of *mentions*, which imply different types of linguistic units:
- Noun phrases (NPs):
    a) pronouns (including elliptical and clitic pronouns)
    b) full NPs (i.e. with a nominal head)
    c) proper nouns
- Discourse segments
- Sentences

2) Linking mentions that corefer, i.e. refer to the same entity.

Since the AnCora corpora already contain syntactic annotations, mentions are already identified if we profit from the annotated syntactic constituents. However, not all NPs are referential and with a view to distinguishing the referential status of NPs, we enrich their tag with one attribute that specifies this information. In addition, two further attributes add information relevant at the level of mentions.

The second task, the annotation of coreference chains, is the main focus of this manual, since it is not trivial to decide whether two mentions refer to the same entity and, to this end, the AnCora coreference coding scheme distinguishes several tags.

Both the addition of attributes to mentions and the annotation of coreference chains are carried out in the same graphical interface: AnCoraPipe, designed and tuned for the Ancora corpora. It offers a specific tool, Coreference annotator (which can be selected from the menu), which facilitates the annotation of coreference.

3. MENTION ATTRIBUTES

The Coreference annotator tool in AncoraPipe offers four different attributes that the coder can select to include if appropriate for the following syntactic units:

- `sn`
- `relatiu`
- `grup.nom` (only when they are a constituent of a larger conjoined NP)
- `v` (when they contain an incorporated clitic, only in Spanish, see Section 4.1.1)

3.1. ENTITYREF

Given that only referential NPs are to be treated as markables, this attribute serves to capture the referential status of an NP. The coder can tick one of the following three values:

- **"spec"** (specific): The NP mention refers to an entity which is named (i.e. it refers to a NE), although the mention itself cannot be considered an NE.
  Ernest Martínez Izquierdo va ser deixeble de Jesús López Cobos. . . . Ø Ha dirigit les principals formacions espanyoles.
  . . . va ser l'única referència que va fer Sanz a la situació de l'equip.
  Després de l'empat amb la reial, el dirigent blanc va anunciar que . . .

- **"nne"** (non-named entity): The NP mention corresponds to a referential entity, but one which is not as "identifiable" as an NE, and so has no unique name. This value includes impersonal third-person pronouns as well as editorial pronouns. Bare NPs complementing a deverbal noun are also entityref="nne".
  Ø Ha dirigit les principals formacions espanyoles, on Ø ha deixat clara la seva serietat.
  La rutina i els formalismes són l'enemic principal de la música clàssica.
  Una trajectòria sòlida i el seu rigor l'avalen per a futurs reptes.
  L'efecte 2000 era un problema real, encara que tots hem ajudat a magnificar-lo.
  Ø No busqueu tres peus al gat.
  ...s'embrancaran en una lluita, com Ø fan a Lleida amb els moros i cristians.
  Qualsevol orquestra bona ha de ser flexible i polivalent.
  Molts el consideren especialitzat en contemporània.

- **"lex"** (lexicalized): The NP mention has been lexicalized and become part of a fixed phrase. It includes inherent clitics in pronominal verbs.
  Ø No busqueu tres peus al gat.
  donar la raó; estar a disposició de l'organització; d'aquesta manera; sense fer declaracions
  tant l'un com l'altre; un per un
  Seria millor, Ø és clar, que tots visquéssim de manera feliç i saludable.
  Amb el canvi de sistema polític no n' hi haurà prou per resoldre els problemes econòmics.

Table 1. Catalan pronominal verbs.

Clitic	Examples
la / les es ho	*ballar-la, cagar-la, carregar-se-la, dinyar-la, dir-ho tot, doldre's, fer-la, fer-ho, fer-s'ho, fer-se, haver-se-les, jugar-se-la, passar-les magres, passar-(s')ho bé, pirar-se-les, tenir-se-les*
en	*anar-se'n, dir-ne, estar-ne, no haver-n'hi per (a) tant, sortir-se'n, tenir-ne prou amb/de, tornar-se'n*
hi	*caure-hi, dir-hi la seva, entendre-hi, fer-hi, ¡Ja hi som!, jugar-s'hi, mirar-s'hi, pensar-s'hi, sentir-hi, ser-hi, tenir-hi (alguna cosa/res) a fer, tocar-hi, tornar-s'hi, trobar-s'hi, valer-s'hi, veure-(s')hi*

Mentions for which **NO** entityref attribute is added include:

- Attributive NPs (functioning similarly to an adjective). It includes pronouns replacing an adjective, bare NPs complementing a non-deverbal noun, and predicative complements.
  No quiero decir que lo sea, cínico o divertido.
  El título de campeón mundial de boxeo.
  El primer ministro israelí, Ehud Barak, llegó hoy jueves a Lisboa.
  El proyecto prevé la utilización de gas natural como combustible principal.

- Nominal predicates
  L. van Gaal y J .Luis Nuñez fueron el centro de las críticas.
  La hipótesis de la colisión era la más probable.
  Los militantes que hoy tienen una emoción especial.

- Appositive phrases
  Andrés Palop, guardameta del Valencia, estará cuatro meses de baja.

- Frequency or duration references
  Tras dos meses de trabajo en torno al naufragio.
  Andrés Palop estará cuatro meses de baja.

- Negated NPs
  No hay conclusiones definitivas sobre la colisión.
  No se les exige ninguna prueba de capacitación.

- Interrogative pronouns
  Las dudas sobre quien ganará las elecciones.

- Coordinated NPs
  La ayuda que nos han enviado la comunidad internacional y los países involucrados.

3.2. homophoricDD

This optional attribute is included for definite NPs with "entityref=nne" or "entityref=spec" that are self-sufficient definite descriptions (DDs). By *self-sufficient* it is meant non-anaphoric, i.e. they are not interpreted by reference to another element.
TEST: they can be the first mention of an entity in a text.

- Proper-noun-like or generic DDs that refer to something in the cultural context or world view, which might have not been introduced in the text:
  el sol ; l'ira ; l'actualitat ; les dones ; els preus
- DDs that can mention an entitiy for the first time in a text thanks to the situational context:
  El plusmarquista mundial de altura está en condiciones de ganar la medalla de oro en los Juegos de Sydney. Desde su regreso a las pistas, Ø sólo ha ganado una competición.
  La criatura que es va trobar dimarts . . . Els metges mantenen el nadó a la incubadora . . . Ø busquen els pares.
- Complex DDs whose modifiers provide the information required for the NP to be interpreted and thus make them self sufficient. The fact that one of the modifiers is anaphoric does not exclude the NP from being self sufficient.
  La història nord-americana
  Els policies responsables de la brutal pallissa

3.3. TITLE

The mention is part of a newspaper headline or a subheading.
Rato agraeix les gestions socialistes a la Unió Europea.

Millora de la C-244 entre Sant Pere i Capellades. El director general de Carreteres, Antoni Lluch, va afirmar que . . . El pressupost de les obres supera . . .

4. COREFERENCE LINKS

The annotation of coreference chains consists in giving the same ENTITY number to all coreferring mentions. To this end, when a mention refers for the second time to a previously introduced entity, a "New entity..." is created and the first mention as well as coreferent mentions must be assigned the same ENTITY number.

Given that the delimitation of mentions relies on the previous layers of annotation, multiword expressions cannot be split even if they contain embedded mentions. Hence it is not possible for *Andorra* to be annotated for coreference:

- × *El Govern_d'_Andorra . . . el país pirinenc*

Coreferent mentions must obligatory receive a COREFTYPE attribute, and in some cases an additional COREFSUBTYPE attribute. The values they can take are now discussed in detail.

4.1. ident (identity)

The identity relation links referential `sn` or `grup.nom` constituents[3] (i.e. with an ENTITYREF attribute other than "lex") that point to the same discourse entity. Preference is given to the speaker's reference over semantic reference. To check for the existence of a link, apply the transitivity test:

TRANSITIVITY TEST for coreference: if mention D is said to corefer with mention C (1), then this means that the slot occupied by D can be occupied by C with no change in meaning (2), namely mention C can successfully combine with the context of D. The same applies to the rest of mentions in the chain up to A.

(1) *La expansión de la piratería en el Sudeste de Asia puede destruir las economías de la región*$_C$ *. . . El papel vital del transporte marino se ve minado por las bandas internacionales, lo que puede llevar a destruir el progreso económico de las naciones de la región*$_D$

(2) *El papel vital del transporte marino se ve minado por las bandas internacionales, lo que puede llevar a destruir las economías de la región*$_C$

Many combinations of `sn` mentions are possible, as shown by the collection in Table 2. A mention corresponding to a syntactic unit of `grup.nom` type can be coreferentially annotated if it is contained within a larger conjoined `sn` unit, as it is the case of the first mention of *Internet* in:

La central de reserves permetrà contractar directament a través d'[Internet, fax, telèfon] o bé a través de les agencies de viatges, qualsevol oferta turística. Les tres associacions han presentat avui Girona Turística, un portal a Internet en què . . .

SPECIAL ISSUES

4.1.1. Clitic pronouns

Catalan clitics (joined to the verb by a hyphen) are identified at the syntactic level under a `sn` unit. In the case of Spanish, where clitics are completely incorporated into the verb (e.g. *verlo*), the ENTITY number is given to the `v` unit as well as the corresponding ENTITYREF attribute.

The Catalan pronouns *en* and *hi* are marked as coreferential with an NP, although from a syntactic point of view they substitute for a prepositional phrase, a prepositional object or an adjunct.

De la reunió sí que en van transcendir imatges.
El Pere es casa demà amb la Marta. El Pere s'hi casa demà.
En les protestes pel cas de Diallo hi conflueixen opinions vàries.

[3] See 4.1.1. for the cases when the `v` unit is annotated for the clitic it contains.

Table 2. Sample of mentions linked by identity link.

Anchor	coreferent mention	Example	
proper noun	proper noun	*Artur Mas*	*Mas*
	full NP	*la Unesco* *Atapuerca*	*el comitè patrimonial* *tota la serra*
	3rd p. pronoun	*el senyor Tony Blair*	*li*
	1st / 2nd p. pronoun (in quoted speech)	*Capitani*	*"…em…"*
	clitic pronoun	*el Barça*	*guanyar-lo*
	demonstrative pronoun	*l'alcalde Rudoph Giuliani*	*aquest*
	relative pronoun[4]	*Wilt Chamberlain*	*que*
	zero pronoun	*la Unesco*	*Ø*
full NP	proper noun	*la ciutat*	*Lugo*
	full NP (same head)	*la muralla romana* *la lluita de classes* *una carrera que havia de provar qui organitzava el pla més aparatós*	*la muralla* *la lluita de classes* *la carrera*
	full NP (synonym, hypernym or hyponym)	*el jaciment d'Atapuerca* *el sistema d'elecció del rector*	*el jaciment burgalès* *la mesura*
	3rd p. pronoun	*els partits*	*els*
	1st / 2nd p. pronoun (in quoted speech)	*la germana Capitani* *l'alcalde Xavier Ulldemolins*	*"...jo..."* *"...Ø* [*espero*]*..."*
	clitic pronoun	*l'efecte 2000*	*magnificar-lo*
	demonstrative pronoun	*les narcosales*	*aquestes*
	relative pronoun	*un període històric* *les institucions financeres*	*que* *les quals*
	zero pronoun	*els constituents*	*Ø*

4.1.2. Quoted speech

Although first- and second-person pronouns are usually deictic, they become anaphoric when they are part of a quoted speech segment within a larger discourse:

Ha explicat la germana Capitani. "Jo no demanava un favor celestial, un miracle."

Editorial first person plural (zero) pronouns are linked when the community can be identified, either in the form of an explicit plural NP or in the form of an organization name.

En paraules d'un dels directius de l'agència, "Ramón y Cajal ens va deixar tirats".

If impersonal pronouns or editorial first-person plural pronouns appear more than once within quoted speech, they are linked.

L'alcalde va anunciar l'elaboració de noves propostes: "Ø estem preparant un projecte, que Ø anomenarem Àrea Temàtica de la Palmera".

[4] DEs that are the anchor of a relative pronoun always contain the relative clause within themselves, as the relative clause is a modifier of the head noun.

4.1.3. Possessives

For possessive NPs, the reference of the entire NP is taken into account. No relation is marked for the possessor expressed by the determiner/pronoun, as these do not correspond to a syntactic sn.

El Partido Demócrata aprobó hoy ofrecer la candidatura a Albert Gore, que fue aclamado con aplausos ante la convención de su partido.

4.1.4. Embedded NPs

Mentions which are embedded within a larger mention are also candidates to participate in a coreference chain, irrespective of the entity to which the larger mention refers.

[*el primer dia de circulació de* **[***l'euro***]**$_1$ **]**$_2$ **. . .** *l'estrena de l'euro*

$$\frac{1}{2}$$

Follow the Maximality Principle: if the reference of an embedded NP pragmatically coincides with that of the entire NP, then the latter is preferred:

Los Angeles **. . .** *la ciutat de Los Angeles . . .*

× ******Los Angeles* **. . .** *la ciutat de Los Angeles*

la Generalitat **. . .** *el Govern de la Generalitat . . .*

× ******la Generalitat* **. . .** *el Govern de La Generalitat*

The maximal NP principle also applies for constructions of the form *the members of (the set)*, since in principle references to a set coincide with references to the set of members, such as mentions to a football team and mentions to all its players. So only the largest unit is annotated:

Argentina ganó ayer **. . .** *los jugadores de Argentina . . .*

× ******Argentina ganó ayer* **. . .** *los jugadores de Argentina*

4.1.5. Relative pronouns

Relative pronouns corefer with the sn unit that contains them:

[*El palmerar urbà d'Elx, que va ser construït a l'edat mitjana pels àrabs***]**

Amb **[***les institucions financeres sense el concurs de les quals Veneçuela no sortirà del fangar.***]**

Complex relatives *(Catalan cosa que*, Spanish *lo que*) are not separately annotated. Instead, the whole relative phrase is linked with the larger sn node:

Aquest espectacular creixement es traduirà en **[***un augment igual de les operacions de càrrega i descàrrega a la ciutat, cosa que farà necessària una nova regulació d'aquesta activitat.* **]**

Relative pronouns are not coreferenced when the relative phrase is nominalized (the head noun is elliptic).

× *És aquest partit polític el que ha menyspreat* **. . .**

× *Ø És el que a Espanya va pretendre prematurament el sindicat vertical franquista.*

× *La curiositat dels que investiguen els dinosaures*

4.1.6. Split antecedent

Coreference between coordinated mentions is possible:

[*Stanford i Bradley***]** *sostenen que Clovis i el Solutrià són gairebé idèntiques* **. . .** *La teoria de tots dos arqueòlegs tardarà anys a avaluar-se.*

Not always, however, the different singular NPs appear linked by coordination. In case of a split antecedent, a new entity is formed adding each of the subconstituents: entity# + entity# (+...):

David H.P. circulava en un ciclomotor Suzuki Katana per Consell de Cent i Ø va ser envestit per un Citroën XM, matrícula d'Almeria. El testimoni no va poder precisar qui dels dos es va saltar el semàfor en vermell.

The converse is not annotated: mentions that are subentities of an entity introduced earlier in the discourse are not linked, as this implies a link type other than coreference, namely part-of or set-member.

Un partit obert fins al final per les ocasions de gol a les dues porteries . . . El Racing va buscar la porteria contrària.

4.1.7. Generic vs. specific NPs

Generic NPs can enter into identity coreference when used referentially:

Un presumpte frau en la venda de gasoil per estafar en els últims dos anys més de 315.000 litres de gasoil . . . els rellotges dels comptadors dels camions de distribució del gasoil.

Coreference links are annotated at a specific and a generic level, but keeping these two levels distinct. Hence, lexical match between the heads of two referential NPs does not suffice to establish coreference. In the following sentence, no relation is coded between the first specific *un nen de la seva classe* and the second generic *un nen*.

Una nena de 6 anys va morir per un tret que va sortir d'una arma que aguantava un nen de la seva classe . . . la indignació de saber que un nen va poder aconseguir una arma i portar-la a l'escola.

Likewise, type and token distinctions for time-dependent entities are kept separate.

El Teatre de Palamós serà l'escenari diumenge a les set del concert "La tenora més enllà de la cobla". Aquest concert es va crear l'any 2000. D'ençà aquest espectacle ha visitat diferents indrets de Catalunya. El preu de les localitats pel concert de diumenge és 3 euros.

4.1.8. Metonymy

The reference of a word on its own right might idffer from its reference when used within a discourse, as captured by Kripke's (1977) distinction between semantic reference and speaker's reference. Consequently, metonymy accounts for mentions that are used to refer to an entity which is associated with the entity originally denoted by the mention. Metonymy within the same newspaper article is annotated as a case of identity, since despite the rhetorical device both mentions are pragmatically used to refer to the same entity. It is just a matter of how the discourse entity is codified in the text.

La Unesco va ser ahir generosa amb Espanya . . . A més a més de Boí i la Tarraco romana, el comitè patrimonial va alabar els valors del jaciment d'Atapuerca . . .

L'Ajuntament de Mollerussa va aprovar una ordenança que prohibeix la venda de begudes alcohòliques en horari nocturn. Gasolineres, màquines expenedores i locals que no tinguin la categoria de bar no podran vendre alcohol.

× Identity of reference cannot be partial but must be complete. If the total identity of referents cannot be ascertained, no relation is encoded.

Los productores centroamericanos . . . los productores de café de Costa Rica.

. . . un producte nacional brut de 570.100 milions de dòlars ... El PNB recull tota la producció del país, inclosa l'exterior.

4.1.9. Lack of agreement

Coreference relations are encoded whether or not there is number and gender agreement. Cases of disagreement do occur:

La Policia Local d'Igualada va inaugurar ahir l'edifici 092, la nova caserna . . . El PSC va arribar a afirmar que en les instal·lacions s'hi havien invertit 240 milions.

. . . el malestar que ha provocat la visita a Gibraltar d' un membre de la Corona anglesa, el príncep Eduard. "A ningú li agrada" que Ø facin aquests viatges.

4.1.10. Missing title

The AnCoraPipe offers the option of adding a missing title to which an NP at the beginning of the text makes reference. The missing NP can then be given the same ENTITY number as the rest of the chain.

($) ($) ($)
L'extracció s'haurà de fer sota la supervisió de l'Administració i Ø inclourà tota la medul·la espinal i els ganglis de l'arrel dorsal. El decret adapta la norma de la UE que . . .
($) ($)
Part del pont ha estat construït a Espanya, que ocupa el quart lloc en la participació després de Suècia, Dinamarca i Alemanya, i davant d'Holanda, França i el Regne Unit.

4.2. pred (predicative)

This link is not properly coreferential because it covers NPs which function as attributions rather than as referential. However, it is appropriate to link them under a special "predicative" COREFTYPE since these relations contain useful information that can be helpful when training a computational coreference resolution system. It takes an obligatory COREFSUBTYPE attribute with two possible values:

4.2.1. definite

The "definite pred" link is reserved for the following three constructions:

(i) nominal predicates (... *es/son* ...) of the identificative kind. By *identificative* it is meant that these NPs – often but not always introduced with the definite article – identify the anchor by one of its defining properties.
(ii) appositional phrases (... , ... ,) of the identificative kind.
(iii) acronyms.

Table 3 presents an example from each type realised by different forms. The head of each predicative construction is distinguished from the attribute according to the following specificity scale:

proper noun > pronoun > definite NP > indefinite

In cases where the two members are equivalent in specificity, the left-most member is marked as the referent.

Table 3. Sample of mentions linked by predicative link.

Anchor	coreferent mention	Example	
proper noun	nominal predicate	*Ricardo Goizueta*	*el responsable de la divisió de comerç electrònic d'El Corte Inglés*
full NP		*les dones*	*el col·lectiu amb sous més baixos*
zero pronoun		*Ø*	*l'autor del fet*
proper noun	appositional phrase	*Andreu Mas-Colell*	*el conseller d'Universitats*
full NP		*el subdirector del Patrimoni Històric de la Xunta*	*Francisco Castro*
proper noun	acronym	*la Comissió Nacional del Mercat de Valors*	*(CNMV)*

Similar predicative relations but different from the three outlined above are not annotated:

- Proper nouns complementing the previous noun
  l'exfuncionari de la comissió José María Ruiza de la Serna
- Subject or direct object complements
  Bauzá acompanyava Camacho com a assessor jurídic.
  La companyia es denominava Bolsa_Consulting.

4.2.2. indefinite

The "indefinite pred link" marks predicative phrases that, although they are not identificative (as the definite ones are), they point out an outstanding characteristic of the referential entity.

Acusat de conservador, ell va ser un dels hereus de John Coltrane que no van travessar la frontera de la tonalitat.
José Bauzá, germà del secretari del consell de la Comissió Nacional del Mercat de Valors, ...
Bauzá també era assessor de l'HSBC.

4.3. dx (discourse deixis)

A mention enters into a relation of discourse deixis when it corefers with a previous discourse segment. Discourse deixis is defined in syntactic terms as these are clearer for coders. The mentions that more frequently have a clausal or sentential antecedent are:

- The neuter pronouns *ho, això* (Catalan) and *lo, esto, eso* (Spanish).
- Elliptical pronouns.
- Definite or demonstrative NPs that are nominalizations of a previous predicate, e.g. *la proposta*.
- Quasi-pronominal definite descriptions of the kind *la cosa*, *el fet*, *la situació*. They can be replaced by the neuter pronoun *això* or *esto* and are almost semantically empty.

The discourse segment (containing one verb at least) that is given an ENTITY number depends on the encoded syntactic nodes: the node that most closely approximates the relevant discourse segment is chosen according to the following scale: `S > sentence > grup.verb > infinitiu`. The dx relation takes an obligatory COREFSUBTYPE attribute with three possible values:

4.3.1. token

The `sn` mention and the discourse segment represent the same event-token, the two events sharing the spatial and temporal coordinates.

*La primera donant d'aquesta nova etapa va ser una dona d'uns 40 anys, a qui es van extreure el cor, el fetge i els ronyons **. . .** Aquesta intervenció havia creat una gran expectació, i des de fa uns quants dies.*
*La lluita de classes ha quedat clausurada**.** Ø Ha estat una clausura solemne.*

4.3.2. type

The two coreferent expressions share the event-type, i.e. the event is of the same kind but not the same occurrence.

Ø va demanar un esforç per assimilar l'euro amb rapidesa i no deixar-ho per més endavant.
El 1966, la monja va vomitar sang. El fet es va repetir al cap de sis mesos.

4.3.3. prop (proposition)

The coreferent mention refers to a previous string of words as a linguistic object *per se* rather than what is being pointed to in the discourse model.

Tornen a castigar els de sempre, els seguidors, mentre_que els que s' emportaran la plata viuen tan a gust. Ø no ho dic per quedar com un heroi; Ø és el que Ø sento com_a aficionat.

"L'euro té potencial per a una apreciació, basada en el creixement i l'estabilitat de preus interna", van declarar ahir conjuntament els ministres d'Economia i Finances del 11 països integrats en la moneda única i el president del BCE. La declaració institucional va ser emesa pel Consell de l'Euro.

If the entity is first introduced by a discourse segment that extends beyond the sentence level: two, three, four, or up to five sentences, then each sentence is assigned an ENTITY number and they are then grouped (entity#+entity#+...) to form the anchor of the NP. This is usually the case for DDs that function as pro-forms in order to encapsulate or package a stretch of written discourse.

Latinoamérica concluyó hoy su participación en la "Bolsa de Turismo" de Berlín con un balance preliminar un tanto pesimista porque Ø no tuvo la cantidad de visitantes esperada. La competencia de Asia, continente de "moda" en la actualidad para los turistas europeos, los altos precios de los pasajes y la relación dólar-marco alemán, fueron los obstáculos señalados por varios países para impedirles lograr sus objetivos. La escasa presencia de interesados provocó que en algunos puestos el material no se distribuyera por completo. Fernández se mostró optimista con respecto a que la situación mejore.

5. REFERENCES

Carlson, Gregory. 1977. A unified analysis of the English bare plural. *Linguistics and Philosophy*, 1(3):413-456.

Halliday, Michael A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman English Language Series 9, Longman, London.

Kripke, Saul. 1977. Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, 2:255-276.

LDC. 2006. ACE Spanish Annotation Guidelines for Entities – Version 1.6 2006.11.01. http://ldc.upenn.edu/Projects/ACE.

Lyons, Christopher. 1999. *Definiteness*. Cambridge University Press.

Poesio, Massimo. 2000. MATE Dialogue Annotation Guidelines – Coreference. Deliverable D2.1. http://www.ims.uni-stuttgart.de/projekte/mate/mdag.

Poesio, Massimo. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Boston, 154-162.

Webber, Bonnie Lynn. 1978. *A Formal Approach to Discourse Anaphora*. Garland, New York.