ITALIAN CONTENT ANNOTATION BANK (I-CAB): PERSON ENTITIES (V. 1.3)[1]

Emanuele Pianta*, Manuela Speranza*, Bernardo Magnini*,
Valentina Bartalesi Lenzi**, and Rachele Sprugnoli**

** ITC-irst, Povo 38050 (Trento) Italy*
{ pianta | manspera | magnini } @itc.it

*** CELCT, Trento 38100 Italy*
{sprugnoli | bartalesi}@celct.it

Maggio 2007

The content of the document is the result of the discussion which has taken place within the working group of ITC-irst and CELCT composed by Valentina Bartalesi Lenzi, Christian Girardi, Bernardo Magnini, Matteo Negri, Emanuele Pianta, Lorenza Romano, Manuela Speranza, Rachele Sprugnoli.

[1] This work has been supported by the ONTOTEXT (From Text to Knowledge for the Semantic Web) project, funded by the Autonomous Province of Trento under the FUP-2004 research program.

TABLE OF CONTENTS

ABSTRACT

This document reports on the annotation of Person entities for the Italian Content Annotation Bank (I-CAB) being developed at ITC-irst in conjunction with CELCT.

I-CAB is a corpus of Italian news annotated with semantic information at different levels. The first level is represented by temporal expressions, the second level is represented by different types of entities (i.e. person, organizations, locations and geo-political entities), and the third level is represented by relations between entities (e.g. the affiliation relation connecting a person to an organization).

As we intend I-CAB to become a benchmark for various automatic Information Extraction tasks, we followed a policy of reusing already available markup languages. In particular, we adopted the annotation schemes developed for the ACE Entity Detection and Time Expressions Recognition and Normalization tasks for English. We describe the extensions to the ACE guidelines needed to adapt them to the specific morpho-syntactic features of Italian and to include a wider range of entities (e.g. conjunctions), providing a large number of examples.

1. INTRODUCTION

This report presents the annotation of Person entities in the Italian Content Annotation Bank (I-CAB), a corpus of semantically annotated documents for Italian containing annotations of temporal expressions (Lavelli et al. 2005), person entities, organization entities, location entities, geo-political entities and of a number of selected relations among such entities.

Following a policy of reusing already available markup languages, the annotation activity has been carried out adopting the formalisms developed within the American ACE program. However, due to the differences between English and Italian, part of the work has been dedicated to the revision and adaptation to Italian of the annotation guidelines.

The creation of I-CAB is part of the three-year project Ontotext funded by the Autonomous Province of Trento. Ontotext focuses on the study and development of innovative knowledge extraction techniques to produce new or less noisy information to be made available for the Semantic Web. Within the new research area of Ontology-Based Knowledge Extraction, Ontotext addresses three key research aspects: annotating documents with semantic and relational information, providing an adequate degree of interoperability of such relational information, and updating and extending the ontologies used for Semantic Web annotation. The concrete evaluation scenario in which algorithms will be tested with a number of large-scale experiments is the automatic acquisition of information about people from newspaper articles.

This document is structured as follows. Section 2 presents the Entity Detection task (valid for all types of entities) and introduces Callisto, the annotation tool we have chosen, as well as the notational conventions used throughout the document. Section 3 describes the annotation of person entities giving several examples and Section 4 provides details on how we deal with pronouns and adjectives referring to person entities. Some special cases are described in Appendix A, while Appendix B presents data about the inter-annotator agreement. Appendix C shows statistics about the person entities annotated in I-CAB and, finally, Appendix D contains the complete list of the text files making up the corpus.

2. THE ENTITY DETECTION TASK

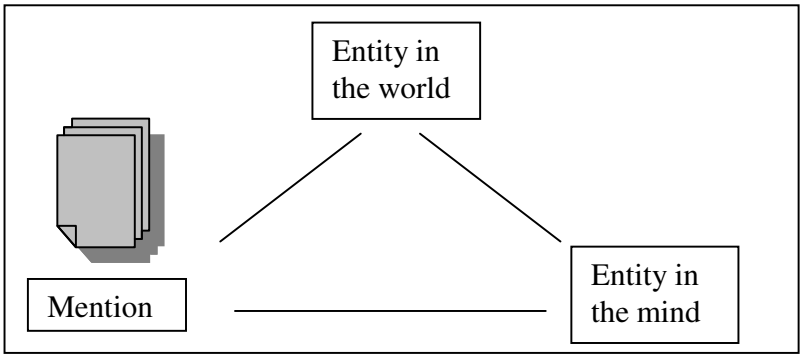
Entity Detection is one of the tasks of the ACE (*Automatic Content Extraction*)[2] program. It "requires that selected types of entities mentioned in the source data be detected, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity" (LDC 2005, p. 4).

The ACE guidelines crucially distinguishes between entities and entity mentions:

- "An entity is an object or set of objects in the world." (LDC 2005, p. 4)
- "A mention is a [textual] reference to an entity. Entities may be referenced in a text by their name, indicated by a common noun or noun phrase, or represented by a pronoun" (LDC 2005, p. 4).

As shown above, the official definition of entity is 'object in the world'. However, in the ACE-LDC guidelines, this term is sometimes used in a way that is not compatible with this definition. Following the ACE-LDC guidelines, for example, the entity *10.000 persone* in *10.000 persone hanno partecipato alla sfilata* (= *10,000 people took part in the parade*), should be classified as underspecified as the exact number can not be quantified. In this case, it makes more sense to say that what is not quantified is actually the mental representation of it (entity in the mind). For the sake of clarity, when we refer to entities in general in this report, if not otherwise specified, we prefer to consider them as mental representations.

Figure 1: Semantic triangle of ENTITIES and Entity MENTIONS



According to the ACE-LDC guidelines, annotators should tag all mentions of each entity within a document; for each mention, they identify the maximal extent of the string that describes the entity and label the head of the mention. Annotators also group co-referring mentions (i.e. all mentions within a text which refer to the same entity).

Mentions can be nested; that is, a mention can contain mentions of other entities or even embedded mentions of the same entity.

Entities are classified from the semantic point of view, so we have different semantic types (e.g. persons, organizations, locations, etc.), subtypes (e.g. persons of subtype individual, group, etc., organizations of subtype sport, commercial, educational, etc.) and reference classes (e.g. specific referential, underspecified, etc.).

[2] http://www.nist.gov/speech/tests/ace/index.htm or http://www.ldc.upenn.edu/Projects/ACE/

Mentions, on the other hand, are classified according to syntactic categories (e.g. proper names, common nouns, pronouns, etc.) and to attributive versus non attributive use. Table 1 presents the complete hierarchy of annotation categories for entities and mentions.

Annotation categories may appear in the report in an abbreviated form (i.e. only the part in upper case). These abbreviations coincide with the labels used in the annotated files, with the only exception of ACETYPE, which becomes simply TYPE, due to the design of the Callisto user interface (see Section 2.5).

Table 1: Annotation Categories

- ENTITY
  - Semantic TYPE
    - Semantic SUBTYPE
  - Reference CLASS
- MENTION
  - ACETYPE
  - LDCTYPE
  - ATtRibutive use

2.1 Semantic TYPES and SUBTYPES of ENTITIES

In the ACE project, seven semantic TYPES of ENTITIES were identified:

- Person: a single individual or a group of humans.
- Organization: corporations, agencies, and other groups of people defined by an established organizational structure.
- Geo-Political Entity: geographical regions defined by political and/or social groups (e.g. a nation, its region, its government, or its people).
- Location: geographical ENTITIES such as geographical areas and landmasses, bodies of water, and geological formations.
- Facility: buildings and other permanent man-made structures and real estate improvements.
- Vehicle: physical devices primarily designed to move an object from one location to another.
- Weapon: physical devices primarily used as instruments for physically harming or destroying other ENTITIES.

In I-CAB we have restricted our annotation to four of the semantic TYPES defined above, while adding the new semantic TYPE Mixed as follows:

- Person (PER)
- Organization (ORG)
- Geo-Political Entity (GPE)
- Location (LOC)
- Mixed (MIX): non-uniform ENTITY groups for which it is impossible to chose a single semantic TYPE (e.g., in *he and his company*, we have a group consisting of a PER and an ORG).

For each semantic TYPE, various SUBTYPES are defined, that provide further semantic information. For instance, for PER, we have the individual, group and indefinite semantic SUBTYPES.

2.2 Reference CLASSES of ENTITIES

Following the ACE-LDC guidelines, each ENTITY (entity in the mind) is assigned a CLASS that describes the kind of reference it makes to something in the world (entity in the world).

Specific Referential (SPC)

An ENTITY is SPC when it refers to a particular, unique object (or set of objects).

[*Ciampi*] *è originario di* [*Livorno*][3]

Notice that an entity can be SPC even if the author or reader is not aware of its name or anchor in the real world.

Non ricordo come si chiama [*il precedente Presidente*]

Generic Referential (GEN)

An ENTITY is GEN when it does not refer to a particular, unique object (or set of objects) but to a general type or class of objects. GEN ENTITIES are typically used in laws and rules.

[*La Camera*] *è eletta a suffragio universale e diretto*

Under-specified Referential (USP)

We reserve the term underspecified for non-specific, non-generic reference. Underspecified references include quantified NP's in modal, future, conditional, hypothetical, negated, uncertain and question contexts: in all these cases the ENTITY/ENTITIES referenced cannot be verified, regardless of the amount of "effort".

Non so [*chi*] *verrà al concerto*

Negatively Quantified (NEG)

An ENTITY is NEG when it has been quantified so as to refer to the empty set of the type of object mentioned.

[*Nessuno stato*] *accetterebbe questo accordo*

2.3 Syntactic categories of Entity MENTIONS: ACETYPES and LDCTYPES

The original ACE annotation scheme, called APF (ACE Program Format), includes four syntactic categories (ACETYPES). In a second phase, in 2004, the Linguistic Data Consortium (LDC) introduced a new annotation scheme called ALF (ACE LDC Format) which adds a number of new syntactic categories (LDCTYPES). We have decided to

[3] In our examples MENTIONS are enclosed in brackets and their head is underlined; when the head is the same as the extent, the underlining is omitted (see Section 2.6) .

follow both indications and thus to annotate each Entity MENTION with an ACETYPE and an LDCTYPE.

The original four ACETYPES are: NAM (proper name), NOM (quantified nominal constructions), PRE (pre-modifier), and PRO (pronoun). For the purpose of adapting the guidelines to Italian, we have transformed PRE into MOD (modifier). In addition, we have added the new ACETYPE MIX (see Section 3.3), so we end up with five ACETYPES (Table 3).

The LDCTYPES have been adopted with a few modifications:

(1) In order to adapt the guidelines to Italian, we have created two separate tags for pre-modifiers (PRE) and post-modifiers (POST), and two new tags, ENCLIT and PROCLIT, respectively for enclitics and proclitics that are attached at the end or the beginning of a word (see Section 4.4).

(2) In addition, we have created a new LDCTYPE, CONJ (see Section 3.4), to annotate conjunctions of ENTITIES (e.g *uomini e donne anziani*; *madre e figlio*). This allows us to mark the co-reference with anaphoric MENTIONS which might follow in the text (e.g. *essi = they, queste persone = these people*, etc.). A CONJ (see Table 2) is a complex MENTION whose head corresponds to the entire MENTION. It also contains nested MENTIONS, in the sense that the different parts of a CONJ are also annotated (e.g. *uomini* and *donne* in the first example, and *madre* and *figlio* in the second, are independent MENTIONS)[4].

Table 2: Some CONJ patterns

x e y	*x and y*
x, y e z	*x, y and z*
tra x e y	*between x and y*
con x e y	*with x and y*
x o y	*either x or y*
x, y o z	*x, y or z*
con x e con y	*with x and with y*
x ma anche y	*x but also y*
x, y ma anche z	*x, y but also y*
della *x* e della *y*	*of the x and of the y*

The adapted lists of the ACETYPES and LDCTYPES used in the annotation of I-CAB are presented in Table 3. The mapping between the set of ACETYPES and LDCTYPES is presented in Tables 5 and 6 (our adaptations in italic).

[4] Please notice that our annotation differs from the ACE-LDC guidelines where conjunctions of entities are not marked explicitly. In the case of *old men and women = uomini e donne anziani* (constructions of conjoined heads that share the same modifiers), they would annotate two simple MENTIONS with the same extent (*old men and women*), the first having *men* as head, the second having *women* as head. In the case of *mother and child* (constructions of conjoined heads without common modifiers), they would only annotate two distinct MENTIONS, *mother* and *child*.

Table 3: ACETYPES and LDCTYPES for the annotation of I-CAB

ACETYPES	LDCTYPES
NAM (proper name) NOM (nominal compound) PRO (pronoun) MOD (modifier) MIX (mixed)	NAM (proper name) NOM (nominal compound) PRO (pronoun) PRE (pre-modifier) POST (post-modifier) BAR (bare) WHQ (question word) HLS (headless) PTV (partitive) ENCLIT (enclitic) PROCLIT (proclitic) APP (appositional construction) ARC (appositional construction with relative clause) CONJ (conjunction)

Table 4: Mapping from ACETYPES to LDCTYPES

ACETYPE	LDCTYPE
NOM	NOM
	BAR
	APP
	ARC
	CONJ
	PTV
NAM	NAM
	APP
	ARC
	CONJ
PRO	PRO
	PROCLIT
	ENCLIT
	WHQ
	HLS
	PTV
	APP
	ARC
	CONJ
MOD	*PRE*
	POST
MIX	*CONJ*

Table 5: Mapping from LDCTYPES to ACETYPES

LDCTYPE	ACETYPE
NOM	NOM
BAR	NOM
NAM	NAM
APP	NOM
	NAM
	PRO
ARC	NOM
	NAM
	PRO
CONJ	NOM
	NAM
	PRO
	MIX
PRO	PRO
ENCLIT	PRO
PROCLIT	PRO
PRE	MOD
POST	MOD
WHQ	PRO
HLS	PRO
PTV	PRO
	NOM

MENTION extent and MENTION head rules

For each MENTION, we record its full *extent*. The *extent* of a MENTION consists of the entire nominal phrase including all modifiers, prepositional phrases and relative clauses (e.g. *Ho incontrato* [*degli uomini* [*che*] *amano gli scacchi*]). In case of ambiguous structures, the extent annotated should be the maximal extent. In case of a discontinuous constituent, the extent goes to the end of the constituent, even if that means including tokens that are not part of the constituent (e.g *Ho incontrato* [*degli uomini, ieri al bar,* [*che*] *amano gli scacchi*]).

In addition, for each simple MENTION, the syntactic *head* is marked (e.g: [*Un altro partecipante importante*]). In most cases, the syntactic head of nominal phrases consists of a single word.

We can have heads composed of more than one word in two cases:

(i) proper names: the whole proper name is considered to be the head of the nominal phrase (i.e. both first and family name);

[*Carlo Azeglio Ciampi*] *è toscano*
[*Il saggio Carlo Azeglio Ciampi*] *è toscano*

(ii) expressions whose meaning has a certain degree of non-compositionality: the whole expression is considered to be the head. When in doubt, annotators refer to a reference dictionary (De Mauro 2000) and annotate the whole expression as head if it is recorded as idiomatic expression.

Giovanni è proprio [*un uccello del malaugurio*]

Appositional constructions (APP, e.g. [*La cantante Madonna*]), appositional constructions with relatives (ARC, e.g. [*La cantante Madonna che è in tour*]), and conjunctions (CONJ, e.g. [*Madonna e Prince*]) are complex constructions where the extent rules for simple MENTIONS are hard to apply. Each complex construction has special extent rules and simple MENTIONS within the extent of complex ones are further annotated. According to the ACE-LDC guidelines it is not necessary to annotate heads of complex constructions. However, the annotation tool we have chosen, i.e. Callisto, requires that every MENTION has a head so we have decided to annotate the whole extent as head.

2.4 Attributive use of Entity MENTIONS

MENTIONS can be used attributively and these uses have to be marked. As a general rule, a MENTION is marked as ATR in the following cases (see Section 3.5):

- when it is a predicate complement (e.g. *la capitale italiana* in *Roma è la capitale italiana*) or an objective complement (e.g. *Little John* in *Lo chiamano Little John*)
- when it occurs in the final part/s of an appositional construction (e.g. *azienda leader nel settore* in *Microsoft, azienda leader nel settore*, see Section 3.4) or in the central part of an ARC (e.g. *avvocato* in *Franco, avvocato, che vive in America*, see Section 3.4);
- when it is preceded by expressions such as *in qualità di*, *come* (= *as*), e.g. *insegnante* in *Luigi lavora come insegnante* (see APP in Section 3.4).

2.5 The annotation tool

For the annotation of I-CAB we have chosen Callisto, a freely distributed annotation tool developed at the MITRE Corporation. It supports linguistic annotation of textual sources for any Unicode-supported language and accepts files encoded as UTF-8, US-ASCII and several other character encodings. Callisto is written in Java, taking advantage of its portability and language support; it has been built with a modular design and utilizes standoff-annotation, allowing for unique tag-set definitions and domain dependent interfaces. Stand-off annotation support allows for many different annotation tasks to be represented. For the annotation of Temporal Expressions we have used the TIMEX2 task, whereas for the annotation of ENTITIES we have used the ACE2004 task.

For each MENTION, Callisto provides two separate slots for the ACETYPES and the LDCTYPES, so we found no problems in annotating each MENTION with both kinds of information. For the annotation of the attributive uses, Callisto provides a specific slot which can be filled with the value ATR or left empty.

2.6 Notational conventions

All our examples are in italics. We have two notational conventions: a short form (which gives only information about the extent and head of MENTIONS) and an extended form.

Short notation:

The MENTION is enclosed in brackets and the head is underlined; when the head is the same as the extent, the underlining is omitted.

[*Un altro partecipante*] *ha testimoniato al processo*

[*Giovanni*] *è sposato da anni*

Extended notation:

MENTIONS are still enclosed in brackets and heads are underlined. For each MENTION, its ACETYPE and LDCTYPE are also provided.

[*Marco*] *è andato al cinema.*

ACETYPE=NAM
LDCTYPE=NAM

If an example contains more than one MENTION (of the same ENTITY or of different ones), we identify each entity and mention with a progressive index: E-1, E-2, etc., for ENTITIES; m-1, m-2, etc., for MENTIONS.

Two mentions referring to the same ENTITY:

[*Marco*]$_{E\text{-}1\ m\text{-}1}$ *è andato al cinema con la* [*propria*] $_{E\text{-}1\ m\text{-}2}$ *macchina.*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=PRO
LDCTYPE=PRO

Two mentions referring to different ENTITIES:

[[*I <u>figli</u> di* [*<u>Marco</u>*]$_{E\text{-}1\ m\text{-}1}$]$_{E\text{-}\ 2\ m\text{-}1}$ *sono andati al cinema.*

E-1	m-1	ACETYPE=NAM
		LDCTYPE=NAM
E-2	m-1	ACETYPE=NOM
		LDCTYPE=NOM

In some examples we add values of ENTITY attributes (e.g SUBTYPES and CLASSES); as far as attributive use is concerned, we only specify when ATR=YES.

[*Marco*] $_{E\text{-}1\ m\text{-}1}$ *è* [*un <u>ingegnere</u>*] $_{E\text{-}1\ m\text{-}2}$

E-1 (PER-Indiv.) (SPC)	m-1	ACETYPE=NAM
		LDCTYPE=NAM
	m-2	ACETYPE=NOM
		LDCTYPE=NOM
		ATR=YES

3. GUIDELINES FOR THE ANNOTATION OF PERSON ENTITIES

In this Section we describe annotation guidelines for Person ENTITIES.

3.1 Semantic SUBTYPES of Person ENTITIES

PER-Individual: when the ENTITY refers to a single person.

[*Quella <u>ragazza</u>*]$_{E\text{-}1\ m\text{-}1}$ *si chiama* [*Francesca*]$_{E\text{-}1\text{-}m\text{-}2}$
E-1 (PER-Indiv.) m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=NAM
LDCTYPE=NAM
ATR=YES

[*Ciampi*]$_{E\text{-}1\ m\text{-}1}$ *è nato nel 1920*
E-1 (PER-Indiv.) m-1 ACETYPE=NAM
LDCTYPE=NAM

PER-Group: when the ENTITY refers to more than one person. This includes family names and ethnic and religious groups that do not have a formal organization unifying them.

[*Quei <u>bambini</u>*]$_{E\text{-}1\text{-}m\text{-}1}$ *sono disubbidienti*
E-1 (PER-Group) m-1 ACETYPE=NOM
LDCTYPE=NOM

More examples:
[*Gli <u>avvocati</u>*] *non lavorano gratis*
[*I <u>Rossi</u>*] *sono originari di Pisa*
[*La mia <u>famiglia</u>*] *abita in centro*
[*Gli <u>arabi</u>*] *parlano una lingua appartenente alla famiglia semitica*
[*I <u>Cristiani</u>*] *professano una religione monoteista*

PER-Indefinite: when it is not possible to judge from the context whether the ENTITY refers to one or more than one person.

Non sappiamo ancora [*chi*]$_{E\text{-}1\text{-}m\text{-}1}$ *l'abbia rubato*
E-1 (PER-Indef.) m-1 ACETYPE=PRO
LDCTYPE=WHQ

3.2 Reference CLASSES

Reference CLASSES describe the kind of reference each ENTITY makes to something in the world.

Specific Referential (SPC)

An ENTITY is SPC when it refers to a particular, unique object (or set of objects), whether or not the author or reader is aware of the name of the ENTITY or its anchor in the real world. For example, in the sentence *Ho visto Francesca passeggiare con un bambino*, both *Francesca* and *un bambino* are to be annotated as SPC (in the first case, the author and the reader are aware of the name of the ENTITY, in the second they are not).

Generic Referential (GEN)

An ENTITY is GEN when it does not refer to a particular, unique object (or set of objects) but a general type or class of objects. GEN ENTITIES are typically used in laws and rules.

[*Il presidente della Repubblica*] *deve avere più di 40 anni*
[*Gli avvocati*] *non lavorano gratis*

Co-reference between GEN ENTITIES is generally admitted:

Scrive libri per [*i bambini*]$_{E\text{-}1\ m\text{-}1}$ *e da anni si occupa della* [*loro*]$_{E\text{-}1\ m\text{-}2}$ *educazione*

E-1	m-1	ACETYPE=NOM
		LDCTYPE=NOM
E-1	m-2	ACETYPE=PRO
		LDCTYPE=PRO

Under-specified Referential (USP)

It is a non-specific, non-generic reference. It includes:

a) quantified NP's in modal, future, conditional, hypothetical, negated, uncertain, question contexts (in all cases the ENTITY/ENTITIES referenced cannot be verified, regardless of the amount of "effort");
   Non so [*quante persone*] *verranno al corteo.*
b) imprecise quantifications;
   [*Tutti*] *sanno quando ci sarà il corteo.*
c) MENTIONS of a large number of ENTITIES where the actual members of the set are not identifiable and the number used is an estimate;
   [*Oltre 10.000 persone*] *hanno partecipato al corteo.*
d) impersonal and passive pronoun *si;*
   [*Si*] *dice che cadrà molta neve* (impersonal, "*si*" means "people")
   [*Si*] *vende carne* (passive)
e) NPs that the annotator cannot classify.

USP reference of type (a), (c), and (e) above all admit co-reference between each other. Co-reference does not occur, usually, between USP reference of type (b) and (d), with only one exception: co-reference is admitted when the imprecise quantification or the

impersonal pronoun “si” co-refer with some pronominal element in the same clause. For example, in the sentences: ***ci si*** *può capire anche senza la guerra* / ***people*** *can understand* ***each other*** *without making war* and ***tutti si*** *potrebbero capire anche senza la guerra* / ***everybody*** *could understand* ***each other*** *without making war*.

[*Tutti*]$_{E\text{-}1\ m\text{-}1}$ [*si*]$_{E\text{-}1\ m\text{-}2}$ *potrebbero capire anche senza la guerra*

E-1 (USP)	m-1	ACETYPE=PRO
		LDCTYPE=PRO
	m-2	ACETYPE=PRO
		LDCTYPE=PRO

Negatively Quantified (NEG)

An ENTITY belongs to the referential CLASS NEG when it has been quantified so as to refer to the empty set of the type of object mentioned.

[*Nessun avvocato di questo studio*] $_{E\text{-}1\ m\text{-}1}$ *ha più di quaranta anni*

E-1 (NEG)	m-1	ACETYPE=NOM
		LDCTYPE=NOM

NEG ENTITIES can not co-refer:

[*Nessuno*] $_{E\text{-}1\ m\text{-}1}$ *ha chiamato,* [*nessuno*] $_{E\text{-}2\ m\text{-}1}$ *ha risposto*

E-1 (NEG)	m-1	ACETYPE=PRO
		LDCTYPE=PRO
E-2 (NEG)	m-1	ACETYPE=PRO
		LDCTYPE=PRO

The only exception is when the NEG ENTITY co-refers with some pronominal element in the same clause.

[*Nessuno*] $_{E\text{-}1\ m\text{-}1}$ [*si*] $_{E\text{-}1\ m\text{-}2}$ *è ferito nell’incidente*

E-1 (NEG)	m-1	ACETYPE=PRO
		LDCTYPE=PRO
	m-2	ACETYPE=PRO
		LDCTYPE=PRO

Mapping between reference CLASSES and semantic SUBTYPES

SPC ENTITIES can be of any semantic SUBTYPE. NEG ENTITIES have semantic SUBTYPE indefinite, as an empty set has no number, whereas all GEN ENTITIES have SUBTYPE group (see Table 7). As for USP ENTITIES, we distinguish between imprecise quantifications, estimates, impersonal and reflexive pronoun ‘si’, and quantified NP’s in modal, future, conditional, hypothetical, negated, uncertain, question contexts (see Table 7).

Table 7: Mapping between reference CLASSES and semantic SUBTYPES

Reference CLASS		**Semantic SUBTYPE**
SPC		any
GEN		PER-Group
NEG		PER-Indefinite
USP	a) modal, future, etc., context	any
	b) imprecise quantifications	PER-Group
	c)estimates	PER-Group
	d) impersonal and reflexive 'si'	PER-Indefinite

Here are some examples of CLASS-SUBTYPE combinations related to the pronoun *chi*:

[*Chi partecipa all'incontro*] *ha diritto di votare* (GEN-Group)
[*Chi ha votato contro*] *lo ha fatto per le ragioni più disparate* (SPC-Group)
Non riesco a immaginare [*chi*] *possa aver votato contro!* (USP-Indefinite, because one or more people may have voted against)

3.3 MENTION ACETYPES (adapted for I-CAB)

NAM (Names): proper nouns and nicknames.

[*Napolitano*] *è di origine campana*
[*Pinturicchio*] *sta giocando bene*

NOM (Quantified Nominal Constructions): nouns quantified with determiners, quantifiers, or possessives.

[*Il presidente dell'azienda*] *non è nel suo ufficio*
[*Il mio vicino*] *è partito ieri*

PRO (Pronouns): all pronouns and headless MENTIONS have ACETYPE PRO.

[*Loro*] *non conoscono l'inglese*
[*Molti*] *non lo sanno*
[*Qualcuno*] *verrà*
[[*Suo*] $_{E\text{-}2\ m\text{-}1}$ *figlio*] $_{E\text{-}1\ m\text{-}1}$ *è nato nel 2000*

E-1	m-1	ACETYPE=NOM
		LDCTYPE=NOM
E-1	m-2	ACETYPE=PRO
		LDCTYPE=PRO

MOD (Modifiers): as Italian, unlike English, does not admit nouns in modifier position (cfr. the office director versus l'ufficio del direttore), this ACETYPE does not apply for person ENTITIES.

MIX (Mixed MENTIONS): it is used for conjunctions of ENTITIES when the two ENTITIES are not of the same ACETYPE. For example, when one is NAM and the other is NOM, as in the following example (E-3 m-1):

[[*Marco*]$_{E\text{-}1\ m\text{-}1}$ *e* [*alcuni* <u>*amici*</u>]$_{E\text{-}\ 2\ m\text{-}2}$]$_{E\text{-}3\ m\text{-}3}$ *sono andati al cinema.*

E-1 m-1 TYPE=NAM
LDCTYPE=NAM
E-2 m-1 TYPE=NOM
LDCTYPE=NOM
E-3 m-1 TYPE=MIX
LDCTYPE=CONJ

3.4 MENTION LDCTYPES (adapted for I-CAB)

NAM (Names): proper nouns and nicknames.

[*Laura*]$_{E\text{-}1\ m\text{-}1}$ *vive a Roma*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM

[*Il* <u>*Pupone*</u>]$_{E\text{-}1\ m\text{-}1}$ *gioca sempre nella Roma*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM

NOM (Quantified Nominal Constructions): nouns quantified with determiners, quantifiers, or possessives.

[*Il* <u>*presidente*</u> *dell'associazione*]$_{E\text{-}1\ m\text{-}1}$ *ha parlato al congresso*

E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM

HLS (Headless MENTIONS): constructions in which the nominal head is not explicitly expressed. Following the ACE convention, we assign as head the rightmost modifier, i.e. the one which falls directly before the spot where the head would be.

This is the case of superlative adjectives (when the noun they modify is elided), percentages, and numerals used as pronouns (notice that they all have ACETYPE PRO).

[*Il 30* <u>*%*</u>]$_{E\text{-}1\ m\text{-}1}$ *è biondo*

E-1 m-1 ACETYPE=PRO
LDCTYPE=HLS

[*Due*] $_{E\text{-}1\ m\text{-}1}$ *sono europei e* [*tre*] $_{E\text{-}2\ m\text{-}1}$ *sono asiatici*

E-1 m-1 ACETYPE=PRO
LDCTYPE=HLS
E-2 m-1 ACETYPE=PRO
LDCTYPE=HLS

[*Il più <u>forte</u>*]$_{E\text{-}1\ m\text{-}1}$ *vincerà*

E-1 m-1 ACETYPE=PRO
LDCTYPE=HLS

Also indefinite pronouns having the same form as their corresponding adjectives have LDCTYPE HLS. See for instance *molti* (=*many*), *alcuni* (=*some*), *tutti* (=*everyone*), *altri* (=*other*), *tutti gli altri* (=*everyone else*), *quanti* (=*how many*), *entrambi* (=*both*), *gli stessi* (=*the same*), *tutt'e due* (=*both*).

On the other hand, indefinite pronouns having different form than their corresponding adjective have LDCTYPE PRO. See for instance *nessuno* (=*nobody*) as pronoun and *nessun* (=*any*) as adjective.

[*Nessuno*]$_{E\text{-}1\ m\text{-}1}$ *mi sa aiutare*

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

WHQ (WH-Words): interrogative and relative pronouns

[*La <u>ragazza</u>* [*che*]$_{E\text{-}1\ m\text{-}2}$ *vedi dalla finestra*] $_{E\text{-}1\ m\text{-}1}$

E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=PRO
LDCTYPE=WHQ

[*<u>Quelli</u>* [*che*] $_{E\text{-}1\ m\text{-}2}$ *credono*]$_{E\text{-}1\ m\text{-}1}$*agli UFO si riuniscono periodicamente*

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO
m-2 ACETYPE=PRO
LDCTYPE=WHQ

The annotation of the correlative pronoun "chi" (=*those who*) includes the pronoun and also the subordinate clause:

Con grossa delusione di [*<u>chi</u> pensava di farcela*]$_{E\text{-}1\ m\text{-}1}$

E-1 m-1 ACETYPE=PRO
LDCTYPE=WHQ

The annotation of the interrogative pronoun "chi" includes only the pronoun:

[*Chi*] $_{E\text{-}1m\text{-}1}$ *è stato a compiere l'omicidio?*

E-1 m-1 ACETYPE=PRO
LDCTYPE=WHQ

More examples:

[*Chi*]$_{E\text{-}1\ m\text{-}2\ ATR}$ *è per lei* [*Ciampi*]$_{E\text{-}1\ m\text{-}1}$*?* [*Un grande presidente*]$_{E\text{-}1\ m\text{-}3\ ATR}$*?*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=PRO
LDCTYPE=WHQ
ATR=YES
m-3 ACETYPE=NOM
LDCTYPE=NOM
ATR=YES

[*Chi*]$_{E\text{-}1\ m\text{-}1}$ *ruba?* [*Gente comune*]$_{E\text{-}1\ m\text{-}2}$.

E-1 (*USP*) m-1 ACETYPE=PRO
LDCTYPE=WHQ
m-2 ACETYPE=NOM
LDCTYPE=PRO

ENCLIT (Enclitics): enclitics that are attached at the end of a verb ([*incontrarlo*]*/to meet him*) or of the adverb *ecco* ([*Eccolo*]*!/Here he is!*). We take all the word as extent and the prefix as head.

Potresti [*incontrarlo*]$_{E\text{-}1\ m\text{-}1}$ *domani*

E-1 m-1 ACETYPE=PRO
LDCTYPE=ENCLIT

PROCLIT (Proclitics): proclitics that precede the main verb and are attached to another word. We take all the word as extent and only the prefix as head.

[*Aladino*]$_{E\text{-}1\ m\text{-}1}$ *non è il vero nome,* [*glielo*]$_{E\text{-}1\ m\text{-}2}$ *hanno appiccicato*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=PRO
LDCTYPE=PROCLIT

Notice that in Italian proclitics can also occur as free morphemes. In this case they are tagged as ACETYPE PRO and LDCTYPE PRO.

Non [*lo*] $_{E\text{-}1\ m\text{-}1}$ *vedo da sei mesi*

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

PRO (Pronouns): all remaining pronouns

Among others:

a) personal pronouns:

[*Lei*]$_{E\text{-}1\ m\text{-}1}$ *non sa cosa fare*

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

b) possessive pronouns:

Mi piace di più [*il* <u>*tuo*</u>] $_{E\text{-}1\ m\text{-}1}$

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

c) indefinite pronouns whose form differ from the corresponding adjective, such as pronoun *nessuno* (=*nobody*) – adj. *nessun/nessuna* (=*no*); pron. *qualcuno* (=*someone*)– adj. *qualche* (=*some*). Other examples are *ciascuno* (=*everyone*), *quest'ultimo* (=*the latter*), *qualcun altro* (=*someone else*), *quelli/e* (=*those*), *questi/e* (=*these*):

Stasera arriva [*qualcuno*] $_{E\text{-}1\ m\text{-}1}$

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

d) uno (when it means "a fellow, a man"):

Ho visto [*uno*] $_{E\text{-}1\ m\text{-}1}$ *attraversare col rosso!*

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO

BAR (Bare Nominal MENTIONS): unquantified nominal constructions without pre-modifiers and articles. They can be singular or plural, and are very common in appositional constructions.

Un incontro di [*laici*] $_{E\text{-}1\ m\text{-}1}$ *sul tema della fede*

E-1 m-1 ACETYPE=NOM
LDCTYPE=BAR

[*Poliziotto*] $_{E\text{-}1\ m\text{-}1}$ *ucciso sulla strada*

E-1 m-1 ACETYPE=NOM
LDCTYPE=BAR

PTV (Partitive Constructions): they have two elements, the part and the whole. The first element (the part) quantifies over the second element (the whole).

[<u>*alcuni*</u> [*dei* <u>*soci*</u>]$_{E\text{-}2\ m\text{-}1}$]$_{E\text{-}1\ m\text{-}1}$ *sono in riunione*

E-1 (SPC) m-1 ACETYPE=PRO
LDCTYPE=PTV
E-2 (SPC) m-1 ACETYPE=NOM
LDCTYPE=NOM

NB: In Italian, *parte* (= *part*) and *maggioranza (= majority*) are nouns so they are tagged as NOM PTV, whereas, their English correspondent are tagged as PRO PTV in the ACE-LDC corpus.

[<u>*parte*</u> [*dei* <u>*consiglieri*</u>]$_{E\text{-}2\ m\text{-}1}$]$_{E\text{-}1\ m\text{-}1}$ *si è astenuta*

E-1 (SPC) m-1 ACETYPE=NOM
LDCTYPE=PTV
E-2 (SPC) m-2 ACETYPE=NOM
LDCTYPE=NOM

[*la* <u>*maggioranza*</u> [*dei* <u>*deputati*</u>] $_{E\text{-}2\ m\text{-}1}$] $_{E\text{-}1\ m\text{-}1}$ *era favorevole*

E-1 (SPC)	m-1	ACETYPE=NOM
		LDCTYPE=PTV
E-2 (SPC)	m-2	ACETYPE=NOM
		LDCTYPE=NOM

Group nouns (*gruppo*, *famiglia*, *equipe*) can occurr in pseudo-partitive expressions (e.g. *un gruppo di* meaning *un gruppo formato da*). In this case we don't have a partitive construction and the group expression (pseudo-part) co-refers with the pseudo-whole expression.

[*un* <u>*gruppo*</u> *di* [<u>*bambini*</u>]$_{E\text{-}1\ m\text{-}2}$] $_{E\text{-}1\ m\text{-}1}$ *gioca a calcio*

E-1 (SPC)	m-1	ACETYPE=NOM
		LDCTYPE=NOM
	m-2	ACETYPE=NOM
		LDCTYPE=BAR

APP (Appositional Constructions): they consist of a pivot nominal phrase and one or more appositions.

We consider as appositional constructions any combination of various types of contiguous co-referring noun phrases[5]:

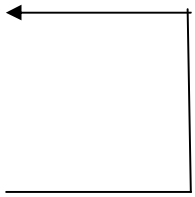
- Noun – noun: [*l'avvocato, grande professionista*], *lavora tantissimo*
- Noun – proper name: [*la cantante Madonna*] *è in tour in Italia*
- Noun – pronoun: [*il mio avvocato, quello che ti ho presentato*], *è in vacanza*
- Proper name – noun: [*Mario, il muratore*], *non è venuto stamani*
- Proper name – pronoun: [*Giovanni, il più forte*], *ha vinto di nuovo*
- Pronoun – noun: [*lui, il tassista*], *è in sciopero*
- Pronoun – proper name: [*uno dei giocatori più forti, Del Piero*], *gioca nella Juve*
- Pronoun – pronoun: [*lei, quella dell'altra volta*], *è ancora in ritardo*

In most of the cases, the second part of the appositional construction (and similarly also the third, the forth, etc.) is ATR (see Section 3.5).

The ACETYPE of any APP is the same as the ACETYPE of the pivot of the APP (generally, the first part).

[[*Sophia Loren*]$_{E\text{-}1\ m\text{-}1}$ [<u>*attrice*</u> *italiana*]$_{E\text{-}1\ m\text{-}2}$]$_{E\text{-}1\ m\text{-}3}$ *è famosa in tutto il mondo*

E-1	m-1	ACETYPE=NAM
		LDCTYPE=NAM
	m-2	ACETYPE=NOM
		LDCTYPE=BAR
		ATR=YES
	m-3	ACETYPE=NAM
		LDCTYPE=APP



NB: the extent of the head is the whole MENTION

[5] Contiguous noun phrases can be separated by a comma, but not by colons, parenthesis, dashes or similar punctuation marks (see Appendix A, "How to deal with dashes, colons and brackets").

We consider titles and honorifics preceding proper names as part of an appositional construction, differing from the LDC guidelines in which titles and honorifics are treated as pre-modifiers:

[[*Papa*]$_{E\text{-}1m\text{-}1}$ [*Giovanni Paolo II*]$_{E\text{-}1\ m\text{-}2}$]$_{E\text{-}1\ m\text{-}3}$ *ha viaggiato molto*

E-1 m-1 ACETYPE=NOM
LDCTYPE=BAR
m-2 ACETYPE=NAM
LDCTYPE=NAM
ATR=YES
m-3 ACETYPE=NOM
LDCTYPE=APP

NB: the extent of the head is the whole MENTION

Constructions in which expressions such as *in qualità di* and *come* (= *as*) are used to connect two nouns, are *not* annotated as APP. Still, MENTIONS introduced by these expressions are ATR.

[*Luigi Rossi*] $_{E\text{-}1\ m\text{-}1}$ *in qualità di* [*produttore*] $_{E\text{-}1\ m\text{-}2}$ *ha realizzato molti film*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES

[*Lei*] $_{E\text{-}1\ m\text{-}1}$ *è brava come* [*mamma*] $_{E\text{-}1\ m\text{-}2}$

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES

On the other hand, if the expression *in qualità di*, *come* is implicit, we annotate an appositional construction:

Una nuova Lazio con [[*Toni*] $_{E\text{-}1\ m\text{-}1}$ [*punta centrale*] $_{E\text{-}1\ m\text{-}2}$] $_{E\text{-}1\ m\text{-}3}$

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES
m-3 ACETYPE=NAM
LDCTYPE=APP

ARC (Complex Constructions taking a Relative Clause): an appositional construction with an adjacent relative clause (WHQ). The relative clause must refer to the initial, referential (SPC) MENTION of the ENTITY rather than the latter, attributive (ATR) MENTION(S) of the ENTITY. Each sub-part of an ARC-construction is tagged.

General rules:

1. the second part of an ARC is attributive

2. the ACETYPE of an ARC is the same as the ACETYPE of the pivot
3. the pivot, its apposition/s and the WHQ MENTION co-refer

[[*l'ex* <u>*direttore*</u>]$_{E\text{-}1\ m\text{-}1}$, [*Rossi*] $_{E\text{-}1\ m\text{-}2}$, [*che*] $_{E\text{-}1m\text{-}3}$ *faceva spesso tardi*] $_{E\text{-}1\ m\text{-}4}$ *è stato licenziato*

E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=NAM
LDCTYPE=NAM
ATR=YES
m-3 ACETYPE=PRO
LDCTYPE=WHQ
m-4 ACETYPE=NOM
LDCTYPE=ARC

CONJ (Conjoined constructions): all conjunctions of ENTITIES. The head coincides with the MENTION extent.

[*I bellissimi* [*Marco*]$_{E\text{-}1\ m\text{-}1}$ *e* [*Giorgio*]$_{E\text{-}2\ m\text{-}1}$]$_{E\text{-}3m\text{-}1}$ *fanno i modelli*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-2 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-3 m-1 ACETYPE=NAM
LDCTYPE=CONJ

Sono arrivati [[*15* <u>*uomini*</u>]$_{E\text{-}1\ m\text{-}1}$ *e* [*20* <u>*donne*</u>]$_{E\text{-}2\ m\text{-}1}$]$_{E\text{-}3\ m\text{-}1}$ *nel campo profughi*

E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
E-2 m-1 ACETYPE=NOM
LDCTYPE=NOM
E-3 m-1 ACETYPE=NOM
LDCTYPE=CONJ

[[*Marco*]$_{E\text{-}1\ m\text{-}1}$ *e* [*Giorgio*]$_{E\text{-}2\ m\text{-}1}$ [*che*]$_{E\text{-}3\ m\text{-}2}$ *sono belli*]$_{E\text{-}3\ m\text{-}1}$ *guadagnano molto*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-2 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-3 m-1 ACETYPE=NAM
LDCTYPE=CONJ
m-2 ACETYPE=PRO
LDCTYPE=WHQ

3.5 Attributive Use

Example of MENTIONS of Person ENTITIES used attributively:

- predicate complements:

[*Franco Rossi*]$_{E\text{-}1\ m\text{-}1}$ *è* [*un* <u>*avvocato*</u>]$_{E\text{-}1\ m\text{-}2}$

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=NOM
ATR=YES

- object complements (Ital. complemento predicativo dell'oggetto): a complement that is used to predicate a description of a direct object.

[*Lo*]$_{E\text{-}1\ m\text{-}1}$ *chiamano* [*Little John*]$_{E\text{-}1\ m\text{-}2}$

E-1 m-1 ACETYPE=PRO
LDCTYPE=PRO
m-2 ACETYPE=NAM
LDCTYPE=NAM
ATR=YES

- in appositional constructions:

[[*Franco Rossi*]$_{E\text{-}1\ m\text{-}1}$,[*l'*<u>*avvocato*</u>]$_{E\text{-}1\ m\text{-}2}$]$_{E\text{-}1\ m\text{-}3}$, *si è trasferito di recente*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=NOM
ATR=YES
m-3 ACETYPE=NAM
LDCTYPE=APP (the extent of the head is the whole MENTION)

- in ARC constructions:

[[*Rossi*]$_{E\text{-}1\ m\text{-}1}$, [*avvocato*] $_{E\text{-}1\ m\text{-}2}$, [*che*]$_{E\text{-}1\ m\text{-}3}$ *vive in America*]$_{E\text{-}1\ m\text{-}4}$ *ha quattro bambini*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES
m-3 ACETYPE=PRO
LDCTYPE=WHQ
m-4 ACETYPE=NAM
LDCTYPE=ARC

- preceded by expressions such as in *qualità di*, *come* (= *as*):

[*Giovanna*] $_{E\text{-}1\ m\text{-}1}$ *lavora come* [*insegnante*] $_{E\text{-}1\ m\text{-}2}$

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES

4. PRONOUNS AND ADJECTIVES

4.1 Personal pronouns

They are all ACETYPE = PRO and LDCTYPE = PRO.

[*Io*] *lavoro a Milano con* [*lui*] *da tre anni*
[*L*]*'ho incontrato ieri al cinema*
[*Mi*] *piacciono le pesche*

4.2 Possessive adjectives and pronouns

Unlike in English, in Italian possessive adjectives and possessive pronouns have the same form.

They are both annotated with ACETYPE and LDCTYPE PRO.

[*Maria*]$_{E\text{-}1\ m\text{-}1}$ *è partita con la* [*sua*]$_{E\text{-}1\ m\text{-}2}$ *macchina.*

E-1	m-1	ACETYPE=NAM
		LDCTYPE=NAM
	m-2	ACETYPE=PRO
		LDCTYPE=PRO

NB: "a sua volta" is considered an idiomatic expression so the possessive is not annotated.

4.3 Indefinite pronouns

An indefinite pronoun is a pronoun referring to an identifiable but not specified person or thing. Examples of English indefinite pronouns are *any*, *none*, *some*, etc. When they refer to people, they are annotated with ACETYPE = PRO but, for the LDCTYPE, we have two options (see Section 2.4):

1. LDCTYPE = PRO. Pronouns whose form differs from the corresponding adjective:
   a) *alcuno, chiunque, ciascuno, nessuno, ognuno, qualcun altro, qualcuno, quelli/e, quest'ultimo, questi/e*
   b) *uno* (when it means "a fellow, a man"): e.g. Ho *visto* [*uno*] *attraversare col rosso!*

2. LDCTYPE = HLS. Indefinite pronouns having the same form as their corresponding adjectives: *alcuni, altri, entrambi, gli stessi, molti, quanti, tutt'e due*, *tutti, tutti gli altri*.

4.4 Clitics

Clitics referring to people are always annotated with ACETYPE PRO. As far as the LDCTYPE is concerned, annotation varies depending on whether they are attached to another word (verb, pronoun or adverb *ecco*) or else separate:

- if they are separate, they have LDCTYPE = PRO
- if they are attached they have
    - LDCTYPE = ENCLIT, when they follow the verb ([*incontrarlo*]*/to meet him*) or the adverb *ecco* ([*Eccolo*]*!/Here he is!*)
    - LDCTYPE = PROCLIT, when they precede the verb ([*glielo*] *dirò/I will tell it to him*)

Sono andato a [*trovarlo*]$_{E\text{-}1\ m\text{-}1}$ *due mesi fa*
E-1 (SPC) m-1 ACETYPE=PRO
LDCTYPE=ENCLIT

[*Gliene*]$_{E\text{-}1\ m\text{-}1}$ *parlerò domani*
E-1 (SPC) m-1 ACETYPE=PRO
LDCTYPE=PROCLIT

[*Li*]$_{E\text{-}1\ m\text{-}1}$ *ho incontrati ieri a teatro.*
E-1 (SPC) m-1 ACETYPE=PRO
LDCTYPE=PRO

Please notice that with clitics attached to another word, the extent of the MENTION includes both the word and the pronoun, and has the pronoun as head.

When used with the past participle (without auxiliary verb), the infinitive, the gerund and the imperative, clitics are always attached to the verb.

4.5 Reflexive pronouns.

Reflexive pronouns can be of four different types:

1. **truly reflexive**, "riflessivi propri": the object of the action is the same as the subject

   [*Mi*]$_{E\text{-}1\ m\text{-}1}$ *sono ferito in montagna*
   E-1 m-1 ACETYPE=PRO
   LDCTYPE=PRO

2. **reciprocal**, "riflessivi reciproci": express mutual action or relationship among the referents of a plural subject, i.e. *each other* in English

   [*Quei due ragazzi*]$_{E\text{-}1\ m\text{-}1}$ [*si*]$_{E\text{-}1\ m\text{-}2}$ *odiano*
   E-1 m-1 ACETYPE=NOM
   LDCTYPE=NOM
   m-2 ACETYPE=PRO
   LDCTYPE=PRO

3. **benefactive**, "riflessivi benefattivi": the focus refers to the person or thing an action is being done for

   [*Mi*] *sono mangiato una mela*
   [*Gli*] *ho mangiato tutte le patatine*

   Sono andato a [*lavarmi*] $_{E\text{-}1\ m\text{-}1}$ *le mani*
   E-1 m-1 ACETYPE=PRO
   LDCTYPE=ENCLIT

4. **pseudo-reflexive**, "riflessivi impropri": occur with intransitive pronominal verbs *Mi sono mosso troppo tardi* ("*muoversi*" is an intrans. pronominal verb)

We annotate all reflexive pronouns with the only exception of the pseudo-reflexive pronouns.

4.6 The impersonal and the passive pronoun "*si*"

In Italian we can have three different clitics which have the same form *si*:

1) a reflexive pronoun (see the previous Section);
2) an impersonal pronoun (in this case it corresponds to *one*, *you*, *we*, or *they* in English)
   [*si*] *dice che* = la gente dice che
3) the so called "si passivante": the *si* pronoun occurs in proclitic position within a clause featuring a transitive active verb with a plural patient realized as post-verbal subject
   Da qui [*si*] *vedono le montagne*

In these cases "si" is always an ENTITY of the USP reference class.

[*si*]$_{E\text{-}1\ m\text{-}1}$ *dice che sia molto intellligente*
E-1 (USP) m-1 ACETYPE=PRO
LDCTYPE=PRO

4.7 Synthesis of the possible interpretations of *si*

For each possible interpretation of the pronoun *si*, Table 8 shows whether it is annotated or not in I-CAB and Table 9 reports some examples.

Table 8: Annotation of the different types of si

Truly reflexive pronoun	YES
Reciprocal pronoun	YES
Impersonal pronoun	YES
Passive *si*	YES
Benefactive pronoun	YES
Pseudo-reflexive	NO

Table 9: Examples of annotations of *si*

Intanto [*si*] *chiedono soldi*	PER (USP) ACETYPE=PRO LDCTYPE=PRO	passive
[*Si*] *può ipotizzare un miglioramento*	PER (USP) ACETYPE=PRO LDCTYPE=PRO	impersonal
È importante che [*si*] *prendano delle decisioni*	PER (USP) ACETYPE=PRO LDCTYPE=PRO	passive
Giovanni [*si*] *è visto un bel film dopo cena*	PER (SPC) ACETYPE=PRO LDCTYPE=PRO	benefactive
[*Si*] *è vista arrivare un mazzo di fiori*	PER (USP) ACETYPE=PRO LDCTYPE=PRO	benefactive

APPENDIX A: Special Cases

Meta-information

In principle, meta-information occurring in the text (e.g. the name of the author) should be annotated as such. However, no metadata field is provided in Callisto, so we have no means for doing it. As a consequence, when the names of the journalists who authored the news stories appear in the text (be it their full name or just their initials), they are simply tagged as person ENTITIES.

Titles of books, CDs and exhibitions

Names of people appearing in book/CD titles, in names of organizations or events, etc., are not annotated, consistently with the criteria adopted for the annotation of temporal expressions[6].

Articulated prepositions

According to the ACE-LDC guidelines, definite and indefinite articles are considered as part of the textual realization of an ENTITY, while prepositions are not. This is problematic for Italian articulated prepositions, where a definite article and a preposition are merged. We have decided that this type of prepositions should be included in the extent of the MENTION, so as to consistently include all the articles.

Person versus other semantic TYPES of ENTITIES

In general, groups of people are annotated as Person ENTITIES. However, there are some particular cases in which they are annotated as Geo-Political ENTITIES or Organization ENTITIES.

a. Person versus Geo-Political ENTITIES

According to the ACE-LDC guidelines, "Geo-Political Entities are composite ENTITIES comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.). All MENTIONS of these four aspects of GPEs will be marked as GPE and co-referenced." (LDC 2005, p. 13)

[*La Francia*] *è* [*una grande nazione*]
Il G8 si riunirà a [*Mosca*]
[*L'Unione Europea*] *ha approvato il progetto*

A MENTION that refers to the entire population of a Geo-Political ENTITY is annotated as GPE, rather than PER.

GPE: [*Gli Italiani*] *amano la pasta*
PER: [*Gli italiani* [*che*] *vivono in America*] *sono tantissimi*

[6] See Lavelli et al., 2005.

NB: *Arabs* (when the word does not refer to the inhabitants of Saudi Arabia) are PER because they do not belong to a unified political structure.

b. Person versus Organization ENTITIES

Organization ENTITIES are "all corporations, agencies, and other groups of people defined by an established organizational structure." (LDC 2005, p. 4)

[*La FIAT*] *nasce nel 1899*
Elena lavora [*alla scuola di Gardolo*]
[*Le scuole*] *chiudono a Giugno*

Political Parties

Political parties are ENTITIES of TYPE ORG. When the text mentions the people belonging to the party (e.g. *diessini*, *leghisti*) instead of the name of the party itself (e.g. *DS*, *Lega Nord*), the entity is classified either as a person or an organization, depending on the context:

- if the text refers to the party as a whole, or to its directives, we have an organization
  [*I leghisti*] *hanno votato contro l'emendamento*
- if the text refers to a specific subset of the people belonging to the party (especially if they behave somehow differently from the others) we have a person ENTITY
  [*I Verdi*] *hanno abbandonato l'aula*
  In this example, the text refers to the people of the green party who were present in the chamber and left it.

Sport teams

Similarly, sport teams can be either organizations or person ENTITIES depending on the context. If the text refers to the team in general, its management or its administration, we have an organization; if instead ir refers to the players, we have a person ENTITY.

ORG: *La maglietta* [*della Juventus*] *è in vendita nei negozi specializzati*
PER: [*La Juventus*] *ha fatto due gol ieri*

Authorities

An authority can be a person or an organization having official power so it is annotated as PER or as ORG depending on the context.

PER: *All'incontro sono intervenute* [*le autorità*]
ORG: *Vogliono aprire il dialogo con* [*le autorità civili della nazione*]

Other specific cases

See Table 10 and 11 for other specific cases.

Table 10: Specific cases of Person

PERSON
(centro)destra, (centro)sinistra
Cast
Centristi
Comunità
cristiani, ebrei, musulmani, buddisti, islamici, induisti
critica (cinematografica)
opposizione, minoranza, maggioranza
resistenza irachena
vertici (of a company)

Table 11: Specific cases of Organization

ORGANIZATION
cattolici
CDA / organi di società
cori, orchestre, bande, accademie
dirigenza
Guardia di Finanza, Carabinieri, il 118, Vigili del Fuoco
forze politiche
giuria
gruppi musicali (REM), band
esecutivo
guardia nazionale irachena
militari fascisti
presidenza europea, presidenza olandese
Procura
resistenza islamica = HAMAS
sindacati, parti sociali

Person versus no ENTITY annotation

Names of people are not annotated if they refer to something which is not a person. For instance, in *la legge Bossi-Fini* or *il governo Prodi*, the names do not identify persons anymore (i.e. the relation with the person has been lost), but they are used as the names of the law or the government itself.

On the other hand, when a preposition is used, names of people are annotated as PER:

La legge di [[*Bossi*] *e* [*Fini*]]
Il governo di [*Prodi*]

In some cases it is difficult to decide whether an ENTITY should be annotated as PER or whether it should not be annotated at all. In these cases, annotators refer to the definition provided in the dictionary (De Mauro 2000): if the *genus* in the definition refers to people, we annotate it as PER, otherwise, it is not annotated at all.

Following the definitions contained in the dictionary, it has been decided to annotate *corteo*, *soccorsi*, *presenze*, *casta*, and *categoria* as PER and not to annotate *volanti della polizia*, *assemblea dei soci* and *tavolo delle trattative*. See the following definitions:

YES Corteo: "gruppo di persone che sfilano nel corso di una manifestazione pubblica: *un c. di manifestanti dimostrava in piazza*, *un c. di studenti*, *di lavoratori*, *un c. militare*, *di protesta*" (De Mauro 2000)

NO Assemblea: "riunione spec. numerosa per discutere questioni importanti di interesse comune: *indire*, *convocare*, *organizzare*, *sciogliere un'a.*; *l'a. di fabbrica*" (De Mauro 2000)

How to deal with dashes, colons and brackets

Dashes, colons and brackets may be relevant for deciding the extent of MENTIONS. This is particular important to decide when a sequence of nominal phrases constitutes two different mentions or an unique appositional construction. The following example illustrate some of the guidelines we followed.

[*La ragazza*] $_{E\text{-}1\ m\text{-}1}$ *(20 anni) è sua sorella*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM

[*La ragazza*] $_{E\text{-}1\ m\text{-}1}$ *(20 anni), molto carina, è sua sorella*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM

[*La ragazza (20 anni)* [*che*] $_{E\text{-}1\ m\text{-}2}$ *vedi alla finestra*] $_{E\text{-}1\ m\text{-}1}$ *è sua sorella*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=PRO
LDCTYPE=WHQ

[*Quella ragazza*] $_{E\text{-}1\ m\text{-}1}$ ([*che*] $_{E\text{-}1\ m\text{-}2}$ *ha 20 anni) è alla finestra da ore*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=PRO
LDCTYPE=WHQ

[*Quella ragazza*] $_{E\text{-}1\ m\text{-}1}$ - [*che*] $_{E\text{-}1\ m\text{-}2}$ *ha 20 anni - è alla finestra da ore*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=PRO
LDCTYPE=WHQ

[*Due guardiani*] $_{E\text{-}1\ m\text{-}1}$ - [*giovani poliziotti*] $_{E\text{-}1\ m\text{-}2}$ *- erano alla finestra*
E-1 m-1 ACETYPE=NOM
LDCTYPE=NOM
m-2 ACETYPE=NOM
LDCTYPE=NOM
ATR=YES

[*Gina*] $_{E\text{-}1\ m\text{-}1}$ ([*studentessa*] $_{E\text{-}1\ m\text{-}2}$*) è alla finestra da ore*
E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES

[*Gina*] $_{E\text{-}1\ m\text{-}1}$ *(20 anni)* [*studentessa all'università*] $_{E\text{-}1\ m\text{-}2}$ *è di Trento*

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
m-2 ACETYPE=NOM
LDCTYPE=BAR
ATR=YES

[[*Fassino*] $_{E\text{-}1\ m\text{-}1}$ *(Ds),* [*Rutelli*] $_{E\text{-}2\ m\text{-}1}$ *(Margherita) e* [*Fini*] $_{E\text{-}3\ m\text{-}1}$ *(AN)*] $_{E\text{-}4\ m\text{-}1}$

E-1 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-2 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-3 m-1 ACETYPE=NAM
LDCTYPE=NAM
E-4 m-1 ACETYPE=NAM
LDCTYPE=CONJ

Interjections

An interjection is a part of speech that usually has no grammatical connection to the rest of the sentence and simply expresses emotion on the part of the speaker: we do not annotate them even if they contain words that in their base sense refer to a person (or a group of person):

Ragazzi, che roba!

"Circa" and "almeno"

These two adverbs are included in the MENTION extent.

[*Circa 10 persone*] *stavano aspettando in uffcio*
[*Almeno 10 persone*] *hanno espresso parere positivo*

APPENDIX B: Inter-annotator agreement

Inter-annotator agreement has been evaluated on the dual annotation of a subset of ten randomly chosen news stories for a total of 4,657 words.

We have adopted the matching criteria of the ACE 2005 distributed scorer:

- an entity is detected by both annotators if they detect at least a mention of that entity;
- a mention is detected by both annotators if the mutual fractional head overlap is at least 30%;
- the maximum extent difference allowed for mentions to be declared an extent match is 4 characters.

Therefore, if one annotates [Savani e Vujevic sempre meglio] / *Savani and Vujevic always better* as a mention while the other restricts the extent to *Savani* e *Vujevic*, we have agreement in mention detection, but no extent match.

The kappa statistic does not account for nested annotations. As this phenomenon is extremely frequent in the case of PEs, we have chosen to calculate the Dice coefficient instead and limit the use of the kappa statistic to the assignment of attributes.

The Dice coefficient is computed as in [1], where C is the number of common annotations, while A and B are respectively the number of annotations provided by the first and the second annotator.

[1] Dice=2C/(A+B)[7]

Results are as follows:

- the Dice coefficient for person entity detection is 0.906;
- limited to the entities detected by both annotators, the Dice coefficient for mention detection is 0.951;
- limited to the entities detected by both annotators, the kappa statistic is 0.937 for subtype assignment (i.e. Group, Individual or Indefinite) and 0.734 for class assignment (this relatively low value is due to the high prevalence of the SPC class and to some mismatches in the USP and GEN classes);
- limited to the mentions detected by both annotators we have a 3.7% of extent mismatch.

[7] Notice that the Dice coefficient has the same value of the F1 measure computed considering any of the two annotators as the reference.

APPENDIX C: Statistical Data

General statistics

	Training	Test	Total
Number of files	335	190	525
Number of words	113,634	68,930	182,564
Average number of words per file	339	362	347
Number of Person ENTITIES	4,530	2,675	7,205
Number of MENTIONS of Person ENTITIES	10,135	6,164	16,299

Distribution of Person ENTITIES by semantic SUBTYPE

	Training	Test	Total
Individual	2,072 (45.74 %)	1,241 (46.39 %)	3,313 (45.98 %)
Group	2,056 (45.39 %)	1,253 (46.84 %)	3,309 (45.93 %)
Indefinite	402 (8.87 %)	181 (6.77 %)	583 (8.09 %)
TOTAL	4,530	2,675	7,205

Distribution of Person ENTITIES by reference CLASS

	Training	Test	Total
SPC	3,556 (78.50 %)	2,190 (81.87 %)	5,746 (79.75 %)
USP	520 (11.48 %)	264 (9.87 %)	784 (10.88 %)
GEN	429 (9.47 %)	211 (7.89 %)	640 (8.88 %)
NEG	25 (0.55 %)	10 (0.37 %)	35 (0.49 %)
TOTAL	4,530	2,675	7,205

Statistics on values of the ACETYPE attribute

- NAM

	# OCCURRENCES
Training	3,354/10,135 (33.09 %)
Test	1,892/6,164 (30.69 %)

- NOM

	# OCCURRENCES
Training	4,552/10,135 (44.91 %)
Test	2,876/6,164 (46.66 %)

- PRO

	# OCCURRENCES
Training	2,179/10,135 (21.50 %)
Test	1,376/6,164 (22.32 %)

- MIX

	# OCCURRENCES
Training	50/10,135 (0.50 %)
Test	20/6,164 (0.33 %)

Statistics on values of the LDCTYPE attribute

- APP

	# OCCURRENCES
Training	680/10,135 (6.71 %)
Test	382/6,164 (6.20 %)

- CONJ

	# OCCURRENCES
Training	523/10,135 (5.16 %)
Test	301/6,164 (4.88 %)

- PROCLIT + ENCLIT

	# OCCURRENCES
Training	90 (1+89)/10,135 (0.89 %)
Test	70 (0+70)/6,164 (1.14 %)

- NOM

	# OCCURRENCES
Training	2,982/10,135 (29.42 %)
Test	1,870/6,164 (30.34 %)

- PRO

	# OCCURRENCES
Training	1,346/10,135 (13.28 %)
Test	762/6,164 (12.36 %)

- BAR

	# OCCURRENCES
Training	784/10,135 (7.74 %)
Test	564/6,164 (9.15 %)

- ARC

	# OCCURRENCES
Training	49/10,135 (0.48 %)
Test	56/6,164 (0.91 %)

- PTV

	# OCCURRENCES
Training	44/10,135 (0.43 %)
Test	44/6,164 (0.71 %)

- NAM

	# OCCURRENCES
Training	2,934/10,135 (28.95 %)
Test	1,640/6,164 (26.61 %)

- WHQ

	# OCCURRENCES
Training	498/10,135 (4.92 %)
Test	356/6,164 (5.77%)

- HLS

	# OCCURRENCES
Training	205/10,135 (2.02%)
Test	119/6,164 (1.93%)

Number of attributive MENTIONS (ATR="yes")

	# OCCURRENCES
Training	1,279/10,135 (12.62%)
Test	872/6,164 (14.15%)

APPENDIX D: Text Files

Training text files divided by date and category

20040907

ATTUALITÀ - NEWS STORIES

1. adige20040907_id405381.txt
2. adige20040907_id405382.txt
3. adige20040907_id405383.txt
4. adige20040907_id405384.txt
5. adige20040907_id405385.txt
6. adige20040907_id405386.txt
7. adige20040907_id405388.txt
8. adige20040907_id405390.txt
9. adige20040907_id405392.txt
10. adige20040907_id405394.txt
11. adige20040907_id405395.txt
12. adige20040907_id405396.txt
13. adige20040907_id405397.txt
14. adige20040907_id405398.txt
15. adige20040907_id405399.txt

CULTURA - CULTURAL NEWS

1. adige20040907_id405408.txt
2. adige20040907_id405409.txt
3. adige20040907_id405410.txt
4. adige20040907_id405411.txt
5. adige20040907_id405412.txt
6. adige20040907_id405414.txt
7. adige20040907_id405417.txt
8. adige20040907_id405418.txt
9. adige20040907_id405419.txt
10. adige20040907_id405420.txt
11. adige20040907_id405424.txt
12. adige20040907_id405425.txt

ECONOMIA - ECONOMY NEWS

1. adige20040907_id405436.txt
2. adige20040907_id405437.txt
3. adige20040907_id405438.txt
4. adige20040907_id405442.txt
5. adige20040907_id405444.txt
6. adige20040907_id405446.txt
7. adige20040907_id405447.txt
8. adige20040907_id405448.txt

SPORT - SPORTS NEWS

1. adige20040907_id405581.txt
2. adige20040907_id405582.txt
3. adige20040907_id405583.txt
4. adige20040907_id405585.txt
5. adige20040907_id405586.txt
6. adige20040907_id405588.txt
7. adige20040907_id405589.txt
8. adige20040907_id405590.txt
9. adige20040907_id405591.txt
10. adige20040907_id405592.txt
11. adige20040907_id405593.txt
12. adige20040907_id405594.txt
13. adige20040907_id405595.txt
14. adige20040907_id405596.txt
15. adige20040907_id405597.txt
16. adige20040907_id405598.txt
17. adige20040907_id405599.txt
18. adige20040907_id405600.txt
19. adige20040907_id405601.txt

TRENTO - LOCAL NEWS

1. adige20040907_id405602.txt
2. adige20040907_id405603.txt
3. adige20040907_id405604.txt
4. adige20040907_id405605.txt
5. adige20040907_id405606.txt
6. adige20040907_id405607.txt
7. adige20040907_id405608.txt
8. adige20040907_id405609.txt
9. adige20040907_id405610.txt
10. adige20040907_id405611.txt
11. adige20040907_id405612.txt
12. adige20040907_id405613.txt
13. adige20040907_id405615.txt
14. adige20040907_id405616.txt
15. adige20040907_id405617.txt
16. adige20040907_id405618.txt
17. adige20040907_id405619.txt
18. adige20040907_id405620.txt

19. adige20040907_id405621.txt
20. adige20040907_id405622.txt
21. adige20040907_id405623.txt
22. adige20040907_id405624.txt
23. adige20040907_id405625.txt
24. adige20040907_id405626.txt
25. adige20040907_id405627.txt
26. adige20040907_id405630.txt
27. adige20040907_id405631.txt
28. adige20040907_id405632.txt
29. adige20040907_id405633.txt
30. adige20040907_id405634.txt
31. adige20040907_id405635.txt

20040908
ATTUALITÀ - NEWS STORIES
1. adige20040908_id405656.txt
2. adige20040908_id405657.txt
3. adige20040908_id405658.txt
4. adige20040908_id405659.txt
5. adige20040908_id405660.txt
6. adige20040908_id405661.txt
7. adige20040908_id405662.txt
8. adige20040908_id405663.txt
9. adige20040908_id405664.txt
10. adige20040908_id405666.txt
11. adige20040908_id405667.txt
12. adige20040908_id405669.txt
13. adige20040908_id405670.txt
14. adige20040908_id405671.txt
15. adige20040908_id405672.txt
16. adige20040908_id405673.txt

CULTURA - CULTURAL NEWS
1. adige20040908_id405684.txt
2. adige20040908_id405685.txt
3. adige20040908_id405686.txt
4. adige20040908_id405687.txt
5. adige20040908_id405688.txt
6. adige20040908_id405689.txt
7. adige20040908_id405690.txt
8. adige20040908_id405691.txt
9. adige20040908_id405692.txt
10. adige20040908_id405693.txt

ECONOMIA - ECONOMY NEWS
1. adige20040908_id405704.txt
2. adige20040908_id405705.txt
3. adige20040908_id405706.txt
4. adige20040908_id405707.txt
5. adige20040908_id405708.txt
6. adige20040908_id405710.txt
7. adige20040908_id405711.txt
8. adige20040908_id405712.txt
9. adige20040908_id405713.txt
10. adige20040908_id405714.txt

SPORT - SPORTS NEWS
1. adige20040908_id405848.txt
2. adige20040908_id405849.txt
3. adige20040908_id405850.txt
4. adige20040908_id405851.txt
5. adige20040908_id405852.txt
6. adige20040908_id405853.txt
7. adige20040908_id405854.txt
8. adige20040908_id405855.txt
9. adige20040908_id405856.txt
10. adige20040908_id405857.txt
11. adige20040908_id405858.txt
12. adige20040908_id405859.txt
13. adige20040908_id405860.txt
14. adige20040908_id405861.txt
15. adige20040908_id405862.txt
16. adige20040908_id405863.txt
17. adige20040908_id405864.txt
18. adige20040908_id405865.txt
19. adige20040908_id405866.txt
20. adige20040908_id405867.txt
21. adige20040908_id405868.txt
22. adige20040908_id405871.txt
23. adige20040908_id405874.txt
24. adige20040908_id405875.txt
25. adige20040908_id405877.txt
26. adige20040908_id405878.txt
27. adige20040908_id405880.txt

TRENTO - LOCAL NEWS
1. adige20040908_id405881.txt
2. adige20040908_id405882.txt
3. adige20040908_id405883.txt
4. adige20040908_id405884.txt

5. adige20040908_id405885.txt
6. adige20040908_id405887.txt
7. adige20040908_id405888.txt
8. adige20040908_id405889.txt
9. adige20040908_id405890.txt
10. adige20040908_id405894.txt
11. adige20040908_id405896.txt
12. adige20040908_id405897.txt
13. adige20040908_id405898.txt
14. adige20040908_id405900.txt
15. adige20040908_id405901.txt
16. adige20040908_id405902.txt
17. adige20040908_id405903.txt
18. adige20040908_id405904.txt
19. adige20040908_id405906.txt
20. adige20040908_id405907.txt
21. adige20040908_id405908.txt
22. adige20040908_id405910.txt
23. adige20040908_id405911.txt
24. adige20040908_id405914.txt
25. adige20040908_id405915.txt
26. adige20040908_id405916.txt
27. adige20040908_id405917.txt
28. adige20040908_id405918.txt

20041007
ATTUALITÀ - NEWS STORIES
1. adige20041007_id413699.txt
2. adige20041007_id413700.txt
3. adige20041007_id413701.txt
4. adige20041007_id413702.txt
5. adige20041007_id413703.txt
6. adige20041007_id413704.txt
7. adige20041007_id413705.txt
8. adige20041007_id413706.txt
9. adige20041007_id413708.txt
10. adige20041007_id413709.txt
11. adige20041007_id413710.txt
12. adige20041007_id413711.txt

CULTURA - CULTURAL NEWS
1. adige20041007_id413719.txt
2. adige20041007_id413720.txt
3. adige20041007_id413721.txt
4. adige20041007_id413722.txt
5. adige20041007_id413723.txt
6. adige20041007_id413724.txt
7. adige20041007_id413725.txt
8. adige20041007_id413726.txt
9. adige20041007_id413727.txt
10. adige20041007_id413728.txt

ECONOMIA - ECONOMY NEWS
1. adige20041007_id413743.txt
2. adige20041007_id413744.txt
3. adige20041007_id413745.txt
4. adige20041007_id413746.txt
5. adige20041007_id413747.txt
6. adige20041007_id413748.txt
7. adige20041007_id413750.txt
8. adige20041007_id413751.txt

SPORT - SPORTS NEWS
1. adige20041007_id413887.txt
2. adige20041007_id413888.txt
3. adige20041007_id413890.txt
4. adige20041007_id413891.txt
5. adige20041007_id413892.txt
6. adige20041007_id413893.txt
7. adige20041007_id413894.txt
8. adige20041007_id413895.txt
9. adige20041007_id413897.txt
10. adige20041007_id413898.txt
11. adige20041007_id413899.txt
12. adige20041007_id413900.txt
13. adige20041007_id413901.txt
14. adige20041007_id413902.txt
15. adige20041007_id413903.txt
16. adige20041007_id413904.txt
17. adige20041007_id413905.txt
18. adige20041007_id413906.txt

TRENTO - LOCAL NEWS
1. adige20041007_id413916.txt
2. adige20041007_id413917.txt
3. adige20041007_id413918.txt
4. adige20041007_id413919.txt
5. adige20041007_id413920.txt
6. adige20041007_id413921.txt
7. adige20041007_id413922.txt
8. adige20041007_id413923.txt
9. adige20041007_id413924.txt

10. adige20041007_id413925.txt
11. adige20041007_id413926.txt
12. adige20041007_id413927.txt
13. adige20041007_id413928.txt
14. adige20041007_id413929.txt
15. adige20041007_id413930.txt
16. adige20041007_id413931.txt
17. adige20041007_id413932.txt
18. adige20041007_id413933.txt
19. adige20041007_id413934.txt
20. adige20041007_id413935.txt
21. adige20041007_id413936.txt
22. adige20041007_id413937.txt
23. adige20041007_id413938.txt
24. adige20041007_id413939.txt
25. adige20041007_id413940.txt
26. adige20041007_id413941.txt
27. adige20041007_id413942.txt
28. adige20041007_id413943.txt
29. adige20041007_id413944.txt
30. adige20041007_id413945.txt

20041008
ATTUALITÀ - NEWS STORIES
1. adige20041008_id413973.txt
2. adige20041008_id413974.txt
3. adige20041008_id413975.txt
4. adige20041008_id413976.txt
5. adige20041008_id413977.txt
6. adige20041008_id413978.txt
7. adige20041008_id413979.txt
8. adige20041008_id413980.txt
9. adige20041008_id413981.txt
10. adige20041008_id413982.txt
11. adige20041008_id413984.txt
12. adige20041008_id413985.txt
13. adige20041008_id413986.txt
14. adige20041008_id413987.txt

CULTURA - CULTURAL NEWS
1. adige20041008_id413995.txt
2. adige20041008_id413996.txt
3. adige20041008_id413997.txt
4. adige20041008_id413998.txt
5. adige20041008_id413999.txt
6. adige20041008_id414000.txt
7. adige20041008_id414001.txt
8. adige20041008_id414002.txt
9. adige20041008_id414003.txt
10. adige20041008_id414004.txt
11. adige20041008_id414005.txt
12. adige20041008_id414007.txt

ECONOMIA - ECONOMY NEWS
1. adige20041008_id414017.txt
2. adige20041008_id414018.txt
3. adige20041008_id414019.txt
4. adige20041008_id414020.txt
5. adige20041008_id414021.txt
6. adige20041008_id414022.txt
7. adige20041008_id414023.txt
8. adige20041008_id414024.txt
9. adige20041008_id414025.txt

SPORT - SPORTS NEWS
1. adige20041008_id414146.txt
2. adige20041008_id414147.txt
3. adige20041008_id414148.txt
4. adige20041008_id414149.txt
5. adige20041008_id414150.txt
6. adige20041008_id414151.txt
7. adige20041008_id414152.txt
8. adige20041008_id414153.txt
9. adige20041008_id414154.txt
10. adige20041008_id414155.txt
11. adige20041008_id414156.txt
12. adige20041008_id414157.txt
13. adige20041008_id414158.txt
14. adige20041008_id414159.txt
15. adige20041008_id414160.txt
16. adige20041008_id414163.txt

TRENTO - LOCAL NEWS
1. adige20041008_id414176.txt
2. adige20041008_id414177.txt
3. adige20041008_id414178.txt
4. adige20041008_id414179.txt
5. adige20041008_id414180.txt
6. adige20041008_id414181.txt
7. adige20041008_id414182.txt
8. adige20041008_id414183.txt
9. adige20041008_id414184.txt

10. adige20041008_id414185.txt
11. adige20041008_id414186.txt
12. adige20041008_id414187.txt
13. adige20041008_id414188.txt
14. adige20041008_id414189.txt
15. adige20041008_id414190.txt
16. adige20041008_id414191.txt
17. adige20041008_id414192.txt
18. adige20041008_id414193.txt
19. adige20041008_id414194.txt
20. adige20041008_id414195.txt
21. adige20041008_id414196.txt
22. adige20041008_id414197.txt
23. adige20041008_id414198.txt
24. adige20041008_id414199.txt
25. adige20041008_id414206.txt
26. adige20041008_id414207.txt
27. adige20041008_id414208.txt
28. adige20041008_id414209.txt
29. adige20041008_id414210.txt
30. adige20041008_id414211.txt

TOTAL: 335

Test text files divided by date and category

ATTUALITÀ - NEWS STORIES
1. adige20040907_id405400.txt
2. adige20040907_id405401.txt
3. adige20040907_id405402.txt
4. adige20040907_id405403.txt
5. adige20040907_id405404.txt
6. adige20040907_id405405.txt
7. adige20040907_id405406.txt
8. adige20040907_id405407.txt
2. adige20040907_id405572.txt
3. adige20040907_id405573.txt
4. adige20040907_id405574.txt
5. adige20040907_id405575.txt
6. adige20040907_id405576.txt
7. adige20040907_id405577.txt
8. adige20040907_id405578.txt
9. adige20040907_id405579.txt
10. adige20040907_id405580.txt

CULTURA - CULTURAL NEWS
1. adige20040907_id405426.txt
2. adige20040907_id405427.txt
3. adige20040907_id405428.txt
4. adige20040907_id405429.txt
5. adige20040907_id405430.txt
6. adige20040907_id405432.txt
7. adige20040907_id405433.txt
8. adige20040907_id405434.txt

ECONOMIA - ECONOMY NEWS
1. adige20040907_id405450.txt
2. adige20040907_id405451.txt
3. adige20040907_id405452.txt
4. adige20040907_id405453.txt
5. adige20040907_id405455.txt

SPORT - SPORTS NEWS
1. adige20040907_id405571.txt

TRENTO - LOCAL NEWS
1. adige20040907_id405636.txt
2. adige20040907_id405637.txt
3. adige20040907_id405638.txt
4. adige20040907_id405639.txt
5. adige20040907_id405640.txt
6. adige20040907_id405641.txt
7. adige20040907_id405642.txt
8. adige20040907_id405643.txt
9. adige20040907_id405644.txt
10. adige20040907_id405645.txt
11. adige20040907_id405646.txt
12. adige20040907_id405647.txt
13. adige20040907_id405648.txt
14. adige20040907_id405649.txt
15. adige20040907_id405650.txt

20040908
ATTUALITÀ - NEWS STORIES
1. adige20040908_id405674.txt
2. adige20040908_id405675.txt
3. adige20040908_id405676.txt
4. adige20040908_id405677.txt
5. adige20040908_id405678.txt
6. adige20040908_id405679.txt
7. adige20040908_id405680.txt
8. adige20040908_id405681.txt
9. adige20040908_id405683.txt

CULTURA - CULTURAL NEWS
1. adige20040908_id405694.txt
2. adige20040908_id405695.txt
3. adige20040908_id405697.txt
4. adige20040908_id405698.txt
5. adige20040908_id405699.txt
6. adige20040908_id405700.txt
7. adige20040908_id405701.txt
8. adige20040908_id405702.txt

ECONOMIA - ECONOMY NEWS
1. adige20040908_id405715.txt
2. adige20040908_id405716.txt
3. adige20040908_id405717.txt
4. adige20040908_id405719.txt
5. adige20040908_id405722.txt

SPORT - SPORTS NEWS
1. adige20040908_id405833.txt
2. adige20040908_id405834.txt
3. adige20040908_id405836.txt
4. adige20040908_id405837.txt
5. adige20040908_id405838.txt
6. adige20040908_id405839.txt
7. adige20040908_id405840.txt
8. adige20040908_id405841.txt
9. adige20040908_id405842.txt
10. adige20040908_id405843.txt
11. adige20040908_id405844.txt
12. adige20040908_id405845.txt
13. adige20040908_id405846.txt
14. adige20040908_id405847.txt

TRENTO - LOCAL NEWS
1. adige20040908_id405919.txt
2. adige20040908_id405920.txt
3. adige20040908_id405921.txt
4. adige20040908_id405923.txt
5. adige20040908_id405925.txt
6. adige20040908_id405926.txt
7. adige20040908_id405927.txt
8. adige20040908_id405928.txt
9. adige20040908_id405929.txt
10. adige20040908_id405930.txt
11. adige20040908_id405931.txt
12. adige20040908_id405932.txt
13. adige20040908_id405933.txt
14. adige20040908_id405936.txt
15. adige20040908_id405937.txt

20041007
ATTUALITÀ - NEWS STORIES
1. adige20041007_id413712.txt
2. adige20041007_id413713.txt
3. adige20041007_id413714.txt
4. adige20041007_id413715.txt
5. adige20041007_id413717.txt
6. adige20041007_id413718.txt

CULTURA - CULTURAL NEWS
1. adige20041007_id413735.txt
2. adige20041007_id413736.txt
3. adige20041007_id413737.txt
4. adige20041007_id413738.txt
5. adige20041007_id413740.txt
6. adige20041007_id413741.txt

ECONOMIA - ECONOMY NEWS
1. adige20041007_id413752.txt
2. adige20041007_id413753.txt
3. adige20041007_id413754.txt
4. adige20041007_id413755.txt

SPORT - SPORTS NEWS
1. adige20041007_id413907.txt
2. adige20041007_id413908.txt
3. adige20041007_id413909.txt
4. adige20041007_id413910.txt
5. adige20041007_id413911.txt

6. adige20041007_id413912.txt
7. adige20041007_id413913.txt
8. adige20041007_id413914.txt
9. adige20041007_id413915.txt

TRENTO - LOCAL NEWS
1. adige20041007_id413946.txt
2. adige20041007_id413947.txt
3. adige20041007_id413948.txt
4. adige20041007_id413949.txt
5. adige20041007_id413950.txt
6. adige20041007_id413951.txt
7. adige20041007_id413952.txt
8. adige20041007_id413953.txt
9. adige20041007_id413954.txt
10. adige20041007_id413955.txt
11. adige20041007_id413956.txt
12. adige20041007_id413958.txt
13. adige20041007_id413959.txt
14. adige20041007_id413960.txt
15. adige20041007_id413961.txt
16. adige20041007_id413962.txt
17. adige20041007_id413963.txt
18. adige20041007_id413964.txt
19. adige20041007_id413965.txt

20041008
ATTUALITÀ - NEWS STORIES
1. adige20041008_id413988.txt
2. adige20041008_id413989.txt
3. adige20041008_id413990.txt
4. adige20041008_id413991.txt
5. adige20041008_id413992.txt
6. adige20041008_id413993.txt
7. adige20041008_id413994.txt

CULTURA - CULTURAL NEWS
1. adige20041008_id414009.txt
2. adige20041008_id414010.txt
3. adige20041008_id414011.txt
4. adige20041008_id414012.txt
5. adige20041008_id414014.txt
6. adige20041008_id414015.txt

ECONOMIA - ECONOMY NEWS
1. adige20041008_id414026.txt
2. adige20041008_id414027.txt
3. adige20041008_id414029.txt
4. adige20041008_id414030.txt
5. adige20041008_id414031.txt

SPORT - SPORTS NEWS
1. adige20041008_id414164.txt
2. adige20041008_id414165.txt
3. adige20041008_id414166.txt
4. adige20041008_id414167.txt
5. adige20041008_id414168.txt
6. adige20041008_id414169.txt
7. adige20041008_id414171.txt
8. adige20041008_id414173.txt
9. adige20041008_id414174.txt
10. adige20041008_id414175.txt

TRENTO - LOCAL NEWS
1. adige20041008_id414213.txt
2. adige20041008_id414214.txt
3. adige20041008_id414215.txt
4. adige20041008_id414216.txt
5. adige20041008_id414217.txt
6. adige20041008_id414218.txt
7. adige20041008_id414219.txt
8. adige20041008_id414220.txt
9. adige20041008_id414221.txt
10. adige20041008_id414222.txt
11. adige20041008_id414223.txt
12. adige20041008_id414224.txt
13. adige20041008_id414225.txt
14. adige20041008_id414226.txt
15. adige20041008_id414227.txt
16. adige20041008_id414228.txt
17. adige20041008_id414229.txt
18. adige20041008_id414230.txt
19. adige20041008_id414231.txt
20. adige20041008_id414232.txt
21. adige20041008_id414233.txt

TOTAL:190

REFERENCES

- (De Mauro 2000) De Mauro, *Il dizionario della lingua italiana per il terzo millennio*, Torino, Paravia, 2000.
  On-line: http://www.demauroparavia.it/
- (Lavelli et al. 2005) Lavelli, Magnini, Negri, Pianta, Speranza, Sprugnoli, *Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 1.0.). Technical Report* T-0505-12, ITC-irst, Trento, 2005.
  On-line: http://tcc.itc.it/projects/ontotext/Publications/i-cab-v1.pdf
- (LDC 2005) Linguistic Data Consortium, *Automatic Content Extraction English Annotation Guidelines for Entities*, version 5.6.1 2005.05.23.
  On-line: http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf
- (LDC 2004) Linguistic Data Consortium, *Mapping from LDC's Annotation Format to ACE Program Format,* version 1.0, 2004-01-30.
  On-line: http://projects.ldc.upenn.edu/ace/docs/Mapping-ALF-to-APF-v1-0.pdf

WEB SITES

- ACE: http://www.nist.gov/speech/tests/ace/index.htm
  http://www.ldc.upenn.edu/Projects/ACE/
- Callisto: http://callisto.mitre.org
- MITRE: http://www.mitre.org/