Resources for linguistically motivated multilingual anaphora resolution

PhD dissertation

Kepa Joseba Rodríguez

Center for Brain and Mind Sciences (CiMEC)
University of Trento

November 2010

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Kepa Joseba Rodríguez
06.12.2010

Acknowledgements

First of all I would like to thank my tutor and advisor, Prof. Massimo Poesio, for his support and encouragement during this research. Next I wish to thank Cogito / Expert Systems, who funded the PhD studentship. I'm grateful to my colleague Francesca Delogu, who worked with me in the organization of the annotation of the corpus. Our discussions about selected examples of Italian texts were very important for the development of the annotation scheme presented in this thesis. Thanks to Olga Uryupina and Egon Stemle for the technical support given for the extraction and processing of the data. I cannot forgot to thank Yannick Versley for the discussions that we have during my research. I learned a lot from his suggestions and comments. Finally I have to thank the annotators that annotated the data sets and signaled possible inconsistences in the annotation scheme. Their contribution was really helpful to introduce corrections in the annotation instructions and without their work this research would not be possible.

Abstract

An actual trend in the computational linguistics and natural language processing is the implementation of multilingual utilities for different tasks, like information retrival, summarization of documents in different languages or machine translation, tasks in which the resolution of anaphoric references plays a crucial role. This dissertation presents a proposal of annotation scheme for the creation of corpus resources for linguistic based multilingual anaphora resolution. This scheme has been implemented for the annotation of English and Italian data. Inter-annotator agreement studies show that the annotation scheme is relaiable. The annotated corpora have been used for the anaphora resolution task, and the results have been compared with well known corpora. Finally hand annotated linguistic features have been used to help in the anaphora resolution process. The results show that our multilingual annotation scheme proposal has been utilized to produce data useful to build anaphora resolution systems for languages with different grammatical and typological features, like English and Italian.

Contents

1 Introduction **13**
1.1 Outline of the dissertation 15

2 Background **17**
2.1 Linguistic background 18
2.1.1 Context dependence 18
2.1.2 Semantic function of nominal expressions 19
2.1.3 Markables 21
2.2 The MUC annotation scheme 24
2.3 The ACE annotation 28
2.4 The MATE proposal 31
2.5 OntoNotes 33
2.6 Other annotation schemes different languages 34
2.7 Summary 35

3 Annotation proposal **37**
3.1 Mentions and markables 38
3.2 Mention surface attributes 42
3.2.1 Annotation of agreement features 42
3.2.2 MIN_words and MIN_IDs 43
3.2.3 Annotation of language specific features 44
3.3 Grammatical function 46
3.4 Reference and no-reference 48
3.4.1 Expletives 48
3.4.2 Predicates 49
3.4.3 Quantifiers 50
3.4.4 Coordination 51
3.4.5 Idioms 51
3.5 Semantic type 51
3.6 Information status 53
3.7 Anaphoric links 53

3.7.1 Annotation 54
3.7.2 Annotation of related object 56
3.8 Conclusions 56

4 Implementation of the proposal for the annotation of Italian and English data 59
4.1 Extraction of text and markables 60
4.1.1 Extraction of markables for the English data 60
4.1.2 Extraction of the Italian data 60
4.2 The annotation tool 61
4.3 Implementation of the scheme 63
4.3.1 Differences between English and Italian annotation schemes 65
4.4 Annotation process 67
4.4.1 Correction of markable boundaries 67
4.4.2 Agreement features and grammatic role 68
4.4.3 Reference and information status 68
4.4.4 Semantic type 69
4.4.5 Type of antecedent and annotation of anaphoric links . 69
4.4.6 Ambiguity 73
4.4.7 Main difficulties reported in the annotation 74
4.5 Description of the annotated data 76
4.5.1 The ARRAU corpus 76
4.5.2 The LMC corpus 76
4.5.3 Statistics of the corpora 77
4.5.4 Reliability studies 80
4.6 Conclusions 81

5 Use of the corpora for anaphora resolution 83
5.1 Tool and classifiers 84
5.1.1 Preprocessing 84
5.1.2 Feature Extraction 85
5.1.3 Language plugin and language specific components . . 85
5.1.4 Learning Modules 87
5.1.5 Evaluation metrics 87
5.2 Baseline features 87
5.2.1 Baseline results for English 88
5.2.2 Baseline results for Italian 89
5.3 Annotation of MIN_IDs 90
5.4 Use of hand annotated features 92
5.4.1 Use of hand annotated gender 92
5.4.2 Use of hand annotated number 93

5.4.3 Use of hand annotated semantic type 94
5.4.4 Use of the grammatic function 94
5.5 Problematic cases 97
5.6 Conclusions 100

6 Conclusions **101**
6.1 Coverage of the annotation 101
6.2 Linguistic criteria 102
6.3 Usability of the annotated resources 102
6.4 Further work 103

List of Figures

4.1 Annotation interface of MMAX 62
4.2 Fragment of the implementation of the annotation scheme for MMAX 64
4.3 Annotation of multiple antecedent 71
4.4 Annotation of discourse deixis in MMAX 72
4.5 Distribution of the non-referring expressions in the ARRAU WSJ dataset 74
4.6 Distribution of the non-referring expressions in the ARRAU WSJ dataset 77
4.7 Distribution of semantic types of referring expressions in ARRAU WSJ dataset 78
4.8 Distribution of the non-referring expressions in the Wikipedia dataset 79
4.9 Distribution of semantic types of referring expressions in Wikipedia dataset 79

5.1 Architecture of BART 84
5.2 Experiment description of BART 86

List of Tables

5.1 Features used by Soon et al (2001) 88
5.2 Baseline results for ARRAU 88
5.3 Baseline results (F1) for MUC 7, ACE-02 and ARRAU 89
5.4 Baseline results for LMC 90
5.5 Baseline results (F1) for ICAB and LMC 90
5.6 ARRAU: Use of MIN IDs 91
5.7 LMC-System: Use of MIN_IDs 91
5.8 Results for ARRAU. Use of hand annotated gender 92
5.9 Results for LMC-Gold. Use of hand annotated gender 93
5.10 Results for LMC-System. Use of hand annotated gender . . . 93
5.11 Results for ARRAU. Use of hand annotated number 94
5.12 Results for LMC-Gold. Use of hand annotated number 94
5.13 Results for LMC-System. Use of hand annotated number . . . 95
5.14 Results for ARRAU. Use of hand annotated semantic type . . 95
5.15 Results for LMC-Gold. Use of hand annotated semantic type . 96
5.16 Results for LMC-System. Use of hand annotated semantic type 96
5.17 Results for ARRAU. Feature: Antecedent is Subject 97
5.18 Results for LMC with gold markables. Feature: Antecedent is Subject 97
5.19 Results for LMC with system markables. Feature: Antecedent is Subject 98
5.20 Results for ARRAU. Feature: Grammatical function matching 98
5.21 Results for LMC with gold markables. Feature: Grammatical function matching 99
5.22 Results for LMC with system markables. Feature: Grammatical function matching 99
5.23 LMC-System: Use of discontinuous markables 100

Chapter 1

Introduction

An actual trend in Natural Language Processing is the development of multilingual utilities. This interest has increased in the actual context of political, cultural and economic integration of European countries, an scenario in which high quality machine translation utilities, information and content extraction applications for different scenarios like transnational federated digital repositories, etc. are always more necessary in science, industry, communications and public facilities.

Multilingual anaphora resolution is essential for the development of last generation applications like multilingual information extraction, machine translation or text summarization.

(1)
1. **German:** Peter hat Maria seine Blumen zum Gießen gegeben. *Sie* hat *sie* vertrocknen lassen.
2. **English (Babelfish):** Peter gave Maria his flowers for pouring. Then *it* left *it* to dry.
3. **English (Google translate):** Peter gave Mary flowers to his casting. Then *she* let *them* dry up.
4. **English (wanted):** Peter gave Maria his flowers to water. Then *she* let *them* dry out.

Example (1) illustrates the importance of multilingual anaphora resolution for the quality of the output of a machine translation system. The German pronoun `Sie` has different possible values for each morphosyntactic feature. The first occurrence of *sie* corefers with the proper name Maria, and the morphosyntactic features are 3rd person, singular, female. The second occurrence of the pronoun corefers with the noun phrase *seine Blumen*, and the agreement features are 3rd person, plural, female. In this case, especially in

the second occurrence of *sie*, a system has to resolve the anaphoric links to be able to produce an interpretation of the pronoun adequate to be reliably translated.

In the translation provided by Babelfish we can see that the German pronoun *Sie* was translated in both cases by the English pronoun *it* with the agreement features 3rd person, singular, neuter. The translation provided by Google Translate is quite better for the second sentence, but introduced an extra pronoun in the first one.

Initial systems for cross-lingual anaphora resolution (Aone and McKee, 1993) relied considerably in domain and linguistic knowledge, resources that are highly time and resources consuming task, and in which only a limited part of the linguistic components are really multilingual.

The availability of annotated corpora for different languages and the advances in statistical NLP, together with the interest in the development of multilingual NLP solutions has increased in the last years the interest in the research about data driven multilingual anaphora resolution.

(Mitkov and Stys, 1997) proposes an approach for high precision pronoun resolution for English and Polish. This approach resolves anaphora in texts pre-processed with part-of-speech tags. The results of this knowledge-poor approach were improved with help of annotated corpora for both languages (Mitkov and Barbu, 2000).

(Harabagiu and Maiorano, 2000) presents a system trained using English and Romanian data. The system was evaluated using the scorer provided for the MUC (Vilain et al., 1995) and achieved for each language similar performance than state of the art monolingual systems.

(Strube, Rapp, and Müller, 2002) adapted the anaphora resolution algorithm of (Soon, Lim, and Ng, 2001) for German data, what allows to use for German a baseline that has been widely used for the evaluation of English systems.

One of the tasks of the 5th International Workshop on Semantic Evaluation (SemEval 2010) is the evaluation of automatic coreference resolution systems for six languages (Catalan, Dutch, English, German, Italian and

Spanish)[1]. Several systems competed to resolve the task in different languages, like BART for English, Italian and German or SUCRE, TANL-1 and UBIU for all the languages (Recasens et al., 2010).

In this context of increasing interest in multilinguality the compatibility between the annotation of data in different languages becomes always more important to train and evaluate multilingual applications.

Some existing annotation schemes have been applied to data in different languages. For instance the annotation scheme of the MUC (Chinchor, 1997) was used to annotate a parallel corpora for English an Rumanian.

The guidelines of ACE (LDC, 2004) and OntoNotes (Pradhan et al., 2007) have been used for the annotation of different languages, like English, Arabic and Chinese and are a potential resource for the training and evaluation of multilingual anaphora resolution systems.

In this dissertation I propose a linguistically motivated annotation scheme to annotate linguistic features and anaphoric links. This scheme has been applied for the annotation of English and Italian datasets.

1.1 Outline of the dissertation

The dissertation is structured in six chapters as follows. The current chapter introduces the issue of multilinguality and anaphora resolution and presents the topic and outline of the thesis.

Chapter 2 gives a linguistic background of the anaphoric references and functions of nominal phrase, and discusses the properties of the markables. After that I review some of the principal annotation schemes used to annotate corpus for anaphora in English and other languages.

Chapter 3 presents the annotation proposal that will be used to build the annotation schemes for English and Italian. This chapter is one of the main contributions of the present dissertation.

Chapter 4 discusses how the annotation proposal has been implemented for the annotation of English and Italian data, and the techniques used to

[1] `http://stel.ub.edu/semeval2010-coref`

extract markables from the text. Finally I give statistics of the annotated data and provide a reliability study of the annotation.

In chapter 5 I present experiments done with the data (more about that after the next draft of chapter 5 is written).

And finally chapter 6 shows the conclusions.(more about that after the next draft of chapter 6 is written).

Chapter 2

Background

In this chapter I give an overview of the current work in the annotation of anaphoric references in text corpora.

First of all in section 2.1 I summarize some linguistic aspects of the anaphora. After that I describe some annotation schemes for English and other languages.

Section 2.2 discusses the anaphoric annotation of the corpora used for the coreference resolution task at the 6th and 7th Message Understanding Conferences (MUC). The MUC corpora were the first high-scale annotation effort.

Section 2.3 describes the anaphoric annotation used in the Automatic Content Extracion (ACE) evaluation campaign.

Section 2.4 describes the MATE proposal, a meta-scheme conceived with the purpose of server as general guideline in the elaboration of annotation schemes. This proposal was tested in the annotation of the GNOME corpus for English (Poesio, 2004b) and the VENEX for Italian (Poesio et al., 2004)

In section 2.5 I present the annotation of OntoNotes, a corpus that partially implements the guidelines proposed in MATE for the annotation of different languages in different levels, from morphology to pragmatics.

Section 2.6 offers an overview of other annotation efforts in different languages, mostly based in one of the previously mentioned approaches.

2.1 Linguistic background

2.1.1 Context dependence

The interpretation of many expressions depends of the context of interpretation, what in Discourse Representation Theory is known as "Universe of Discourse" (Kamp and Reyle, 1993). This context of interpretation includes the linguistic context, visual context of dialogue participants and in general, the shared knowledge or subset of the world under discussion in the discourse situation.

For the purposes of this dissertation I will use **"anaphoric"** to refer to expressions that can be resolved only in the previous linguistic context, and referring expression to expressions that refer to entities of the real world or to the previous discourse.

There are anaphoric expressions can have different kinds of relations with the piece of context necessary for their interpretation. The relation between anaphora and context that probably has been more intensively researched is the identity. An example of identity is the relation between the pronoun *He* and the piece of context necessary to interpret it, the proper name *John*.

(1) *John* has a new car. **He** is very happy.

But there are cases in which anaphora and the part of the context necessary for the interpretation doesn't refer to the same entity in the real world. One of the more relevant cases is the bridging anaphora (Clark, 1975). There are different kinds of bridging relations, like `set-member`, `part-of`, `set-subset`, `other-anaphora`, `attribute-of`, etc.

(2) I meet *two people* yesterday. **The woman** told me a story.[1]

Other relation in which there is not strictly an identity relation between anaphora and the part of the context necessary for the interpretation is the case in which the anaphoric expression refers to a part of the previous discourse, but not to a extern world entity.

(3) As commented in **the previous section**.

The annotation of anaphoric links will be constraint mostly to nominal expressions, which includes the following categories:

[1] Clark:1975

- **Nominals:** noun phrases that have a noun as head.
- **Proper names**
- **Pronouns:** they can be definite, indefinite, demonstrative or reflexive.

2.1.2 Semantic function of nominal expressions

Nominal expressions can play four types of semantic functions: referring, predicative, quantificational and expletive. Only nominal expressions with a referring function are able to be linked to other nominal expressions, or to serve as antecedent of anaphoric expressions. That makes important the implementation of this distinction in annotation schemes for anaphora.

Referring function

That is the function of noun phrases that introduce new entities in the discourse, or are connected with other referring noun phrases by identity or bridging relations.

Predicative function

Predicative noun phrases express properties of world objects. They don't introduce new entities in the discourse and don't refer to previously mentioned entities.

For instance in example (4) the noun phrase *"a computer scientist"* does not introduce a new entity in the discourse, and it is not in a identity or bridging relation with *"Mary"*.

(4) Mary is a computer scientist.

One of the main structures to introduce predication are the copulative sentences, as in example (4), but there are other verbs like *"be called"* (example (5)) or *"be considered"* (example (6)) that can take predicative nominals as argument.

(5) Agnieszka Skrzypek has been called the New Hope of the Polish jazz vocal scene.

(6) He was considered a good teacher of mathematics.

In the given examples the referring nominal appears in the subject position of the sentence, but as we can see in example (7) and (8) the predicative noun phrase can occupy the subject position too.

(7) The new professor is John Shmidt.

(8) A famous researcher is the new professor.

In the example (7) one can see that to be a new professor is a property of the entity *"John Shmidt"*. In example (8) we have a definite noun phrase, *"the new professor* that have already introduced in the discourse, or that is a part of the common ground of the discourse participants, and a property of them realized by an indefinite noun phrase.

Another syntactic construction that often is used to express predication are the appositive constructions like in the following expressions. The predication can appear in the position of the apposition as in example (11) or in the position of the main noun phrase as in examples (9) and (10).

(9) The asbestos fiber, crocidolite

(10) The Nicaraguan president, Daniel Ortega

(11) Giorgio Napolitano, president of Italy

Quantificational function

Quantificational noun phrases denote relations between a set of objects denoted by the nominal expression and the set of objects denoted by a verbal phrase or affected by a property.

(12) all the boxcars

(13) some of the people

Expletives

Forms like *it* and *there* in English or *ci* in Italian can be used to fill a syntactic function without any semantic content as in examples (14) (15) and (16).

(14) **it** would be simple to create hybrids in all crops.[2]

(15) **There** is a large market out there hungry for hybrid seeds

[2]ARRAU corpus, wsj_0209

(16) Cosicchè ci furono due tendenze, spesso riscontrabili nello stesso territorio[3]

They can be used too as pronouns that can be used to previously introduced entities as in examples (17) (18) and (19).

(17) **it** deactivates the anthers of every flower in the plant.

(18) There is a large market out **there** hungry for hybrid seeds

(19) **Ci** troverà a casa.[4]

2.1.3 Markables

The basic unit for the annotation is the markable. The definition of markable that I use here is the one given in the MATE/Gnome project (Poesio, 2004b) *"the text constituent that realize semantic objects that may enter in anaphoric relations"*.

The given definition implies that the markable should not contain only the head of the noun phrase or a minimal projection of it, because in this case we would lost information about the properties of the semantic objects. In most cases the markable should contain the maximal projection of the noun phrase with all the pre- and postmodifiers, including prepositional phrases and relative sentences. For instance the markables of examples (20) and (21) have the same head, and only the information contained in the post-modifiers allows to detect that they refer to different world entities.

(20) The president of France

(21) The president of Nigeria

Gentilic adjectives

The text constituent mentioned in the definition of markable is mostly a noun phrase. But there are other categories of constituents, as in the case of gentilic adjectives, that can realize the anaphora.

(22) The **Mexican** export product

(23) The export product of **Mexico**.

[3] *So there were two tendencies often found in the same territory*
Live Memories Corpus wp_0010

[4] She/he will meet **us** at home.

(24) The best solution to this **Nigerian** problem, is to split **the country** since people have refused have refused to see beyond ethnic boundaries.

In example (22) the gentilic adjective *Mexican* has the same anaphoric function that the proper name *Mexico* in example (23). In example (24) the noun phrase *the country* needs to be linked to the antecedent *Nigerian* in order to interpret its meaning.

Discontinous markables

The maximal projection of the heads of the nominal expression is not the only criteria to build the markable. There are cases in which the semantic material doesn't appear in a linear continuity. For English and Italian that happens mainly in cases of coordination, when the textual representation of two entities share a part of the information, as in examples (25) and (26).

(25) Bill and Hillary Clinton

(26) red and black cars

(27) black cars and bykes

In example (25) we have two different entities, *"Bill Clinton"* and *"Hilary Clinton"*. In this case both entities share the same string of the family name. If we use only the maximal projections of the heads to build the markable as in example (28) we lost information about the mention *"Bill* that is present in the text. In order to keep this information we have to introduce a **discontinuous markable** as in example (29).

(28) Bill
Hillary Clinton

(29) Bill Clinton
Hillary Clinton

In example (26) we have two set of objects realized with the same string as head (*cars*), and distinguished by different premodifiers. If we dont split the set of cars in the two different sets of cars with a distinguishing property as in example (30), then liks between possible definite noun phrases like *"the red cars"* with their antecedent would not be possible. That motivated the introduction of the **discontinuous markable** red cars.

(30) red cars
black cars

In example (27) we have two set of vehicles, cars and bykes, that share the property of being black. If we build the markable using the maxmal projection of the heads as in example (31), then the annotation losts the information about the colour of the bykes, information that can be used for instance to find the incompatibility with other possible markables like *"green bykes"*. A possible way to solve this problem is the creation of the discontinuous markable *"black bykes"* as proposed in example (32).

(31) black cars
bykes

(32) black cars
black bykes

Discourse deixis

Anaphoric expressions don't refer always to entities in the real world, they can refer also to a part of the discourse, like for instance an event or list of events as in examples (33) and (34) or a part of the discourse as in example (35).

(33) *Farm lending was enacted to correct this problem by providing a reliable flow of lendable funds.*
However, **this** in no way justifies the huge government subsidies and losses on such loans.

(34) Last month, the General Accounting Office reported that *defaults in Federal Housing Administration guarantees were five times as high as previously estimated, and that FHA 's equity fell to minus $ 2.9 billion.*
GAO 's findings are particularly troubling because....

Some times the antecedent of the anaphoric expression is not an event, but just a piece of the discourse, like in the example (35) and (36)

(35) *Which are the semantic properties of indefinite noun phrases?*
Can you repeat **the question**.

(36) **That** was a good argumentation.

The interpretation of the anaphora is not always unequivocal. There are cases in which there are alternative interpretations of the meaning of a markable. In example (37) there are two possible interpretations for the pronoun *it*. The first is that the pronoun refers to *the engine*. A alternative interpretation is that the antecedent of the pronoun is the markable *the boxcar*.

(37) Be careful hooking up **the engine** to **the boxcar** because **it** is faulty

2.2 The MUC annotation scheme

One of the main goals of the Message Understanding Conferences was helping in the definition of relevant components for the Information Extraction task. After the observation that the coreference resolution was a crucial component, it was established as a separate task in the 6th and 7th conferences.

The corpora produced for the 6th and 7th conferences are the first large-scale annotation efforts for coreference resolution. The data sets are based on English news papers articles, mostly about economy (from Wall Street Journal) and airlines (New York Times).

(Hirschman and Chinchor, 1997) presents 4 criteria for the Task definition

1. Support for the MUC information extraction tasks
2. Ability to achieve good (around 95%) inter-annotator agreement
3. Ability to mark text up quickly
4. Desire to create a corpus for research on coreference and discourse phenomena, independent of the MUC extraction task.

The corpus is annotated using inline SGML. The annotation consists of adding the `<COREF>` tag to the NPs that are linked by a identity relation.

(38)
```
<COREF ID="100">Lawson Mardon Group Ltd.</COREF>
said <COREF ID="101" TYPE="IDENT" REF="100">it
</COREF> ...
```

The MUC annotation guidelines give instructions for the annotation of newspaper articles and concentrate on the annotation of identity relations between nominal elements. The coreferring expression has three attributes: `ID` number, `TYPE` (always filled with `IDENT`) and `REF`, whose value is the `ID` of the antecedent. There is an optional attribute, `STATUS` that always takes the

value `OPT` and marks optional links, like predications.

In the annotation guidelines, relations can be established between nouns, noun phrases and pronouns. Pronouns include both, personal (including possessive pronouns) and demonstrative pronouns. Dates, percentages and currency expressions are considered nominal phrases.

The annotation of identity relation includes:

- Bound anaphora

  (39) [Most computational linguists] prefer [their] own parsers

  (40) [Every TV network] reported [its] profits yesterday. [They] plan to release full quarterly statements tomorrow.

- Most cases of appositions

  (41) [Julius Cesar], [the well known emperor]

  This identity of reference is to be represented by a coreference link between the appositional phrase, "the well-known emperor" and the ENTIRE noun phrase, "Julius Caesar, the/a well-known emperor" (example (42):

  (42)
  ```
  <COREF ID="1" MIN="Julius Caesar">Julius Caesar,
  <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT">
  the/a well-known emperor, </COREF></COREF>
  ```

- Predicate nominals, including copulas.

  (43) [Bill Clinton] is [the President of the United States].

- Conjoined noun phrases. The individual noun phrases as well as the coordinated noun phrase are potential markables.

  (44) [[the two Croatians] and [Brown]]

  (45) [[so much intelligence] and [so much love]]

- Functions and values. The most recent value is linked to the function.

(46) [The stock price] fell from [$4.02] to [$3.85];

The annotators here have to establish a link between [$3.85] and [The stock price].

Markables for the annotation are the maximal projections of the noun phrase, which contains all the pre-and post modifiers like non-restrictive relative clauses, prepositional phrases, etc. Each markable is annotated with a `MIN` attribute containing the head of the NP as showed in example (47). That makes possible to align markables of the gold standard with markables produced by the system in the case in which the markable boundaries are not exactly the same. That makes possible the evaluation of results.

(47) `But <COREF ID="42" MIN="planes">military training planes</COREF> make up to ...`

If the head of the markable is a multi-world named entity, the full named entity is part of the MIN, like *Julius Cesar* in example (48)

(48) `<COREF ID="1" MIN="Julius Caesar">Julius Caesar, <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT">the/a well-known emperor, </COREF></COREF>`

All named entities annotated as defined in the MUC 6 Named Entity Task Definition are considered markables. But substrings of these named entities are not markables. For instance, if we have the following markables in a text:

(49) [Equitable of Iowa Cos.].... located in [Iowa]

The two occurrences of Iowa are not marked as coreferring, since the first one is just a substring of a named entity.

In case of coordination we have markables with more than one head. Here the annotation of the `MIN` attribute presents some problems.

In example (44) the `MIN` corresponds to the span "Croatians and Brown" as we can see in example (50)

(50) `<COREF ID="59" MIN="Croatians and Brown"><COREF ID="56" TYPE="IDENT" REF="14" MIN="Croatians">The two Croatians </COREF> and <COREF ID="57" TYPE="IDENT" REF="39">Brown </COREF>`

The same happens in example (45), in which the MIN is "intelligence and so much love" (example (51)).

(51) `<COREF ID="60" MIN="intelligence and so much love">so much intelligence and so much love</COREF>`

In both cases the span of the MIN does not correspond with any linguistic category.

The treatment of the predication and the inclusion of bound anaphora in the annotation were criticized by (van Deemter and Kibble, 2000), who argue that the annotation scheme conflates *"elements of genuine coreference with elements of anaphora and predication in unclear and sometimes contradictory ways"*. The main criticisms are:

- Case of bound anaphora.

  (52) [Every TV network] reported [its] profits.

  Following the argumentation of (van Deemter and Kibble, 2000), if we apply the annotation scheme and we annotate [its] as coreferring with [every TV network] the interpretation of the sentence would be that *Every TV network reported the profits of every TV network.*

- Problem of intensionality.

  (53) [Henry Higgins], who as formerly [sales director of Sudsy Soaps], became [president of Dreamy Detergents].

  If we follow the annotation guidelines, then [Henry Higgins], [sales director of Sudsy Soaps] and [president of Dreamy Detergents] must be part of the same coreference set. Since the proposed definition of coreference is a equivalence relation, then [sales director of Sudsy Soaps] and [president of Dreamy Detergents] must be coreferring, what is wrong.

- Problem of predication

  (54) Higgins was/became the/a president of DD.

  (55) Higgins, once the president of DD, is not a humble university lecturer.

In the examples (54) and (55) the predicative element " the/a president of DD" cannot be changed by the proper name "Higgins" without changing the meaning of the sentence. That means that the relation between the constituents is not the IDENT relation required in the guidelines

The annotation scheme of the MUC corpus has a good coverage for all kinds of noun phrases, but the annotation scheme doesn't povide instructions to annotate bridging links and discourse deixis.

2.3 The ACE annotation

In contrast con the MUC annotation scheme and the proposal of MATE discussed in next section, the ACE annotation scheme constrains the annotation to a number of semantic classes considered more relevant for information extraction. This approach is supported by work on systems using machine learning ((Aone and Bennett, 1995) and (McCarthy and Lehnert, 1995)) that constrained the application only to entities of a certain number of semantic classes, mostly person and organization.

A potential advantage of the focus in the annotation of a constrained number of semantic classes, in which the behavior of the surface features has been properly analyzed, is that the determination of identity relations is usually easier to be determined.

The annotation scheme limits the annotation to identity links between mentions to persons, organizations, locations, geopolitical entities, weapons and vehicles(LDC, 2004) The data is stored using the ACE Pilot Format (APF), a stand off XML annotation format.

In APF format each file records all entities annotated in the document with `<ENTITY>` elements. The children of these elements in the XML tree are the `<ENTITY_MENTION>` elements, one for each mention of the entity in the text. Each mention is annotated with the attribute `TYPE`, which have three possible values:

1. `NAM` for named entities.
2. `NOM` for noun phrases which a common noun as head.
3. `PRO` for pronouns.

Each mention is specified with a EXTENT, that specified the span of characters in the original text and contains the string of the markable, and the HEAD which specifies the span of characters and contains the string of the syntactic head of the noun phrase.

(56)
```
<entity_mention ID="2-5" TYPE="NOM" LDCTYPE="NOM"
LDCATR="TRUE">
<extent>
<charseq START="1621" END="1671">an assistant director at
the Oregon Zoo in Portland</charseq>
</extent>
<head>
<charseq START="1634" END="1641">director</charseq>
</head>
</entity_mention>
```

If the head is a named entity realized by more than one word, the full named entity is the head of the markable as in example (57)

(57)
```
<entity_mention ID="1-2" TYPE="NAM" LDCTYPE="NAM">
<extent>
<charseq START="1573" END="1609">American Zoo and Aquarium
Association</charseq>
</extent>
<head>
<charseq START="1573" END="1609">American Zoo and Aquarium
Association</charseq>
</head>
</entity_mention>
```

One of the issues addressed in the ACE annotation guidelines is the treatment of metonymy as in the following examples:

(58) Russia's opposition to the use of force in **Iraq** is the latest in a series of foreign policy disputes with the United States.[5]

(59) Russia, its economy in chaos, desperately needs the cash and also hopes for big new contracts with **Iraq** when sanctions end.[6]

In the example (58) Iraq refers to its geographic extension, and in a further sentence of the same text (example (59)) Iraq refers to the political and eco-

[5] APW19980213.1337
[6] APW19980213.1337

nomical institutions of the country.

The solution proposed by the ACE guidelines to resolve problems caused by metonymy and ensure a higher consistence in the annotation is the creation of the Geopolitical Entity category, that merges the meaning of the country as a physical place, the institution that governs the country and the inhabitants.

The annotation guidelines were used to annotate English, Chinese and Arabic data and has been broadly used to train and evaluate anaphora resolution systems.

One of the problems commented in section 2.2 about the MUC corpus, the problem of predication, has not been be resolved in the annotation of the ACE corpus. For instance in the example (60) we have the markable *"an Asian power"* as coreferring with *"China"*. If both expression would be coreferring, the substitution of China by "an Asian power" would be possible in the example (61) without any change in the meaning of the sentence, what is not the case.

(60) Today , *China* is **an Asian power** and rightfully so.[7]

(61) China has over 1 billion inhabitants.

There is a similar problem with the appositive constructions. In the annotation of ACE appositions are coreferring with the main noun phrase. For instance in example (62) the markable *deputy prosecutor of the war crimes tribunal* corefers with the full noun phrase, what can cause similar problems that the reported for the copulative sentence of example (60)

(62) Graham Blewitt , deputy prosecutor of the war crimes tribunal[8]

A further problem of the problems of the ACE style annotation is that the annotation doesn't cover all the noun phrases. The coverage seems to be sufficient to be a good resource to evaluate state of the art content extraction systems, but it conditions the use of the data for further tasks, like linguistic research of anaphora and discourse phenomena.

Other phenomena that are not covered by the annotation of the ACE corpus are the bridging links and the discourse deixis.

[7] npaper 9801.219
[8] npaper 9801.139

2.4 The MATE proposal

While the goal of the MUC and ACE annotation schemes was to provide resources for evaluation campaigns, the MATE schema was conceived as a meta-scheme that offers potential tags and categories to be used in the development of annotation schemes for different genres and languages.

The MATE coreference scheme (Mengel et al., 2000) considers as markables all the pronouns, noun phrases and proper names occurring in the discourse.

This proposal for annotation of anaphoric information aims to cover a large domain of application than the MUC coreference scheme, including elements necessary for the annotation of other languages, for the annotation of different kinds of text and task oriented dialogues.

In order to annotate references to the visual situation of the dialogue participants, and the frequent use of deixis in task oriented dialogues, the MATE annotation scheme implements the idea of assigning an `ID` to each object in the visual situation. All these objects are grouped in an Universe. The reference to these objects is represented as a link between the markable and the universe element.

(63)
```
<coref:universe ID="u1">
<coref:ue ID="ue1">Diamond mine</coref:ue>
<coref:ue ID="ue2">Graveyard</coref:ue>
<coref:ue ID="ue3">Fast running creek</coref:ue>
<coref:ue ID="ue4">Fast flowing river</coref:ue>
<coref:ue ID="ue5">Canoes</coref:ue>
</coref:universe>
FOLLOWER: Uh-huh.  Curve round.  To your right.
GIVER: Uh-huh.
FOLLOWER: Right....  Right underneath <coref:de ID="de_50">
the diamond mine.  </coref:de> Where do I stop.
GIVER: Well.......  Do.  Have you got <coref:de ID="de_51">
a graveyard?  </coref:de> Sort of in the middle of the page?
...  On on a level to <coref:de ID="de_52">the c-- ...  er
diamond mine.</coref:de>
<link type=''ident args=''51,ue2''>
<link type=''ident args=''52,ue3''>
```

The MATE scheme introduces an extra tag <seg>[9] for the annotation of empty pronouns in Italian, always occurring in the subject position. The tagged span is the full verbal phrase.

(64)
```
Dov'e' <de ID=''157''>Gianni?</de>
<seg type=''pred'' ID=''158''>e' andato a mangiare</seg>
<coref:link href="coref.xml#id(seg_158)" type="ident">
<coref:anchor href="coref.xml#id(de_157)"/>
</coref:link>
```

Although this approach can work for Italian, it is problematic for other romance languages like Portuguese or Spanish in which the empty pronoun can occur in subject and object position at the same time.

A further use of the <seg> tag is the markup of clitic pronouns incorporated to the verb.

(65)
```
A: Mira, te doy <coref:de ID="de_167">este libro</coref:de>
Conoces a <coref:de ID="de_168">mi suegra?</coref:de>
B: S, claro.
A: Pues <coref:seg ID="seg_169">dáselo</coref:seg> cuando
<coref:de ID="de_170">la</coref:de> veas.
<coref:link href="coref.xml#id(seg_169)" type="obj-ident">
<coref:anchor href="coref.xml#id(de_167)"/>
</coref:link>
<coref:link href="coref.xml#id(seg_169) type="iobj-ident">
<coref:anchor href="coref.xml#id(de_168)"/>
</coref:link>
```

The MATE meta-scheme proposes the annotation of discourse deixis. The span of text of the antecedent is marked using the <seg> tag, and annotated with a TYPE attribute that specifies the type of object introduced, an action, an event or a proposition.

(66)
```
GIVER: Youre sort_of going past stone creek ...  but your
lines curving up past the ...  flat rocks.
FOLLOWER: Right.  Okay.
GIVER: <coref:seg ID="seg_135" type="action"> And then starting
to come down again.  </coref:seg>
FOLLOWER: Got <coref:de ID="de_136">that</coref:de>.
<coref:link href="coref.xml#id(de_136)" type="ident">
```

[9] As proposed by the Text Encoding Initiative (TEI)

```
<coref:anchor href="coref.xml#id(seg_135)"/>
</coref:link>
```

The MATE annotation scheme introduce some improvements to the treatment of bound anaphora and predicative noun phrases of the MUC. Anaphoric links to quantified NPs (like "every TV network" in example (52)) are tagged with the label bound. Predicative NPs in copulative constructions are not annotated. This approach was modified for the annotation of the GNOME corpus (Poesio, 2004b) after the observation that the exclusion of predicative NPs leads to difficulties in the automatic extraction of markables. In GNOME the problem was solved introducing a distinction between referring markables, tagged with `term` and non referring markables, tagged with `pred` for predicatives and `quant` for quantified expressions.

MATE includes a set of relations to annotate non-identity anaphoric relations between noun phrases as proposed in DRAMA annotation scheme ((Passonneau, 1997)). The set of annotated relations were reduced to three relations in the annotation of the GNOME corpus, set membership (`ELEMENT`), subset (`SUBSET`) and generalized possession (`POSS`) that includes ownership and part-of relations.

The MATE scheme was used to annotate an Italian corpus, the VENEX (Poesio et al., 2004), that made possible to test the usability of the scheme for the annotations of clitics.

2.5 OntoNotes

OntoNotes (Pradhan et al., 2007) is a large scale corpus for English, Chinese and Arabic annotated at different levels, from part of speech and morphosyntax to discourse and pragmatics.

The anaphoric annotation in OntoNotes uses an scheme close to the MATE guidelines. The annotation is not constrained to noun phrases, and includes verbs if they corefer with a noun phrase and elliptical pronouns in other languages than English. On the other side there are referring expressions that are not included in the annotation of OntoNotes, like gentilic adjectives and nouns and names in premodifier position.

In order to have a higher agreement abstract and underspecified noun phrases are not annotated, but the meaning of these categories is not defined

in the annotation guidelines.

Markables are annotated with semantic categories. The solution given by the annotation guidelines to the problem of metonymy is the opposite to the one proposed by ACE, distinguishing between metonymous and not metonymous uses.

(67) [South Korea] is in South Asia.

(68) [South Korea] has signed the agreement.

The annotation of OntoNotes distinguishes between the mention of South Korea as a country in example (67) and the metonymous mention of South Korea referring to its government of example (68).

As previously mentioned the annotation of OntoNotes doesn't cover nominal and proper noun premodifiers and gentilic adjectives. Other anaphoric phenomena that have not been covered are the bridging references and the discourse deixis.

2.6 Other annotation schemes different languages

The **English-Romanian** corpus presented in (Harabagiu and Maiorano, 2000) is a collection of texts of the MUC-6 and MUC-7 corpora translated into Romanian and annotated using the guidelines of MUC, producing one of the few parallel corpora annotated with anaphoric information. The goal of the annotation was to evaluate the system `SWIZZLE`, a bi-lingual coreference resolver.

The Dutch **CORREA** corpus (Hendrickx et al., 2008) uses an extended version of the MUC guidelines, introducing the annotation of bridging relations and the distinction between predicative relations and identity.

The **TüBa-D/Z** corpus for German (Hinrichs, Kübler, and Naumann, 2005) is a corpus of news papers articles annotated following guidelines inspirated by the MATE proposal. Here the IDENT relation level of MATE is subdivided in three categories:

1. Anaphoric: link between a pronominal anaphora and its closest antecedent..

2. Cataphoric: links a pronoun to the mention that resolves it, if the mention comes later in the text.

3. Coreferent: links definite descriptions to the previous antecedent.

The annotation guidelines of TüBa-D/Z include the annotation of split antecedents for plural mentions.

The **ARRAU** corpus for English (Poesio and Artstein, 2008) and the **LiveMemories** corpus for Italian (Rodriguez et al., 2010) follow closely the MATE guidelines, including the annotation of discourse deixis and some of the relations of the extended schema. An additional feature in the introduce by these corpora is that the schemes give the possibility to the annotators of annotating ambiguity. In the next chapters I will give a detailed description of the annotation scheme and the datasets.

Ancora corpus (Recasens and Mart, 2009) for Spanish and Catalan is based in the MATE scheme, and adds relation types for contextual descriptions and lists.

The annotation of the Ancora corpus includes eliptical and clitic pronouns, discourse deixis and a distinction between referential and attributive or predicative funcion of the noun phrases.

Very few corpora use ACE like guidelines for the annotation. One of the most relevant efforts is the Italian **ICAB** corpus (Magnini et al., 2006) of news paper articles.

2.7 Summary

In this chapter I started by outlining important linguistic issues concerning the annotation of anaphoric references. I have presented some of the most relevant annotation schemes and discussed how some important aspects that have been covered, like predication, bridging or discourse deixis.

In the next chapter I will present the set of guidelines created for the core annotation of the ARRAU Corpus for English and the Live Memories Corpus for Italian.

Chapter 3

Annotation proposal

In this chapter I present the features of the annotation proposal for the implementation of annotation schemes to produce corpus resources annotated with anaphoric information and the differences in some of the features for the annotation of English and Italian texts. In the text chapter I will show how this scheme has been applied for the annotation of English and Italian data.

In this proposal I take as approach to define coreference a discourse model like the proposed by (Webber, 1979), in which coreference resolution happens at the discourse level and is defined as reference to the same discourse entity.

That is one of the reasons why this approach follows very closely the MATE guidelines (Mengel et al., 2000), because they offer the possibility of having links of different types for different kind of relations. That allows to distinguish between coreference and other kind of anaphoric relations.

Other reason why this proposal follows the basic assumptions of the MATE annotation scheme is that we annotate all mentions to real world objects, not only mentions to entities of a set of semantic types. That helps doing possible the implementation of the scheme to annotate texts in different domains and genres. A second important reason is that a corpus annotated with anaphoric information for all mentions will be more useful for general linguistic research, instead in focus its usability in the development of concrete applications.

An additional aspect of MATE is that it offers a useful framework to annotate language specific phenomena of the Italian language, like the annotation of clitics attached to the verb and empty subjects.

In section 3.1 I discuss the identification of markables and markable boundaries. A relevant part of this discussion consists on the introduction of discontinuous markables in the annotation.

Section 3.2 presents the list of surface attributes to be annotated for each markable, like morphosyntactic features and heads of the noun phrases. A important contribution of this section is the proposal for the annotation of two language specific phenomena of the Italian language: the incorporated clitics and the phonologically non realized subjects.

Section 3.3 gives a list of possible grammatical functions that the markables can have in the sentence.

Section 3.4 explains the distinction between referring and non referring markables and lists the types of non referring expressions. In this section tries to avoid the weakness of the annotation of coreference links in the MUC corpus discussed in section 2.2, separating predicative relations from identity.

In section 3.5 I present the semantic types that we will use to annotate the corpora. It is partially based in the ACE annotation scheme.

Section 3.6 explains how the markables will be tagged with information about their information status.

Section 3.7 deals with the annotation of anaphoric links proposed for the corpora, showing the distinction between coreference links and links for other anaphoric relations.

3.1 Mentions and markables

In this proposal all noun phrases will are considered markables for the annotation. Later we will distinguish whether they are mentions of real world entities or not, and if yes they will be annotated with semantic and anaphoric information.

In opposition with the annotation guidelines of OntoNotes and the previous annotation guidelines of ARRAU, nouns and named entities in premodifier position are considered as markables as in examples (1) and (2) if they are part of a coreference chain.

(1) [A [**Lorillard**] spokeswoman]

(2) [the [**US**] government]

(3) [[**oat**] cereal]

For the identification of markables we distinguish between restrictive and non-restrictive relative sentences. Only the relative pronouns of the non-restrictive relative sentences as in example (4) are considered markables.

(4) [The State Secretary, [**who**] met [last week] [the President [of Egypt]]]

Pronouns that are not phonetically realized are considered markables, and their annotation is explained in next chapter.

(5) [**Cesare Battisti**]... 'e stato un [geografo, politico e irredentista italiano].
[∅] Nacque in [Trentino]...[1]

Although we focus on the annotation of noun phrases, there are cases in which other kind of constituents and word classes corefer to real world entities. That is the case of possessive pronouns (6) and gentilic adjectives (7).

(6) South Korea has opened [**its**] market to foreign cigarettes.

(7) The [**Argentinian**] export product

There are two kind of markables which require a separate treatment: coordinations and appositions.

In case of **coordination**, the full coordination and the coordinated noun phrases are markables for the annotation.

(8) "Brothers and sisters"
`[[Brothers] and [sisters]]`

If both coordinated heads share the same determiner as in example (9), each of the coordinated noun phrases have to share the determiner. That makes necessary the introduction of the **discontinuous markables** discussed in section 2.1.3 of previous chapter.

(9) "The brothers and sisters"
`[The brothers and sisters]`

[1] *Cesare Battisti... was an Italian geographer, politician and irredentist. (He) was born in Trentino...*

```
[The Brothers]
[The][sisters]
```

The introduction of discontinuous markables is a feature that was not present in ARRAU or other of the corpora reviewed in chapter 2.

Another situation in which we have to introduce discontinuous markables is when two coordinated heads share the same modifier as in example (10).

(10) "black dogs and cats"
```
[black dogs]
[black][cats]
```

Another case of discontinuous markables in coordinated NPs happens when the same noun is modified by two different modifiers as in (11).

(11) "male and female workers"
```
[male and female workers]
[male][workers]
[female workers]
```

Examples (10) and (11) show the relevance of the introduction of discontinuous markables for the resolution of anaphoric references. For instance, if we splite the coordination as in example (12) we would lost the information about modification of the noun *"cats"*. That is a potential source of errors in the resolution process. For instance, in later in the text we have the mention *"the white cats"* of example (13), the mentions of example (12) and (13) would be potentially coreferring.

(12) "black dogs and cats"
```
[black dogs]
[cats]
```

(13) "the white cats" `[the white cats]`

The case of example (11) is different. If the coordination is splitted as in (14) the mention *"male workers"* would desapear. If later in the text we have the mention *"the male workers"* of example (15), the link to the antecedent would not be possible.

(14) "male and female workers"
```
[male and female workers]
[female workers]
```

(15) "the male workers"
`[the male workers]`

But it is not always possible the introduction of discontinuous markables. If there is no morphosyntactic disagreement in the original text, the extracted markables must be coherent with the morphosyntactic agreement rules of the language.

(16) "the medical schools of Harvard University and Boston University"
`[the medical schools of [[Harvard University] and [Boston University]]]`
`[[Harvard University] and [Boston University]]`
`[Harvard University]`
`[Boston University]`

In example (16), if we define the markables `[the medical schools of Harvard University]` and `[the medical schools of Boston University]` we would have a morphosyntactic disagreement inside of the markable, because "the medical schools" is a plural, but in the Harvard University or in the Boston University there is presumably only a medical school.

Other problematic case is when there is a quantification over the coordinated noun phrase like in example (17).

(17) Five boys and girls
`[Five boys and girls]`

In example (17) we cannot proceed like in example (9), because we would obtain the wrong markables [Five boys] and [Five][girls]. In this case we will not create separate markables for each of the nominal heads.

In case of **apposition**, we have to determine which of the noun phrases is the mention to an entity and which of them is a predication that adds information about it. For instance, in example (18) the noun phrase "chrysotile" refers to a concrete chemical substance, and it must be identified as a markable. The noun phrase *"the common kind of asbestos, chrysotile, found in most schools and other buildings"* is the predicate that adds extra information. That is the second markable of the apposition.

(18) the common kind of asbestos, chrysotile, found in most schools and other buildings
`[the common kind of [asbestos] , [`**`chrysotile`**`] , found in [[most`

`schools] and [other buildings]]`

If there is a named entity in the noun phrase that contains the apposition, it will be identified as a markable for the annotation.

(19) The president of Italy, Giorgio Napolitano
`[The president of Italy,` **[Giorgio Napolitano]**]

(20) Lorillard Inc. , the unit of New York-based Loews Corp. that makes Kent cigarettes
[[Lorillard Inc.] `, the unit of [New York-based Loews Corp.] that makes [[Kent] cigarettes]]`

3.2 Mention surface attributes

3.2.1 Annotation of agreement features

Gender

The proposed values of this feature for the annotation of the corpora are `male`, `female`, `neuter` and `unspecified`.

The feature `gender` takes the value of the grammatical role of the noun phrase, like for the pronouns "he"/"lui" (`gender:male`), "it" (`gender:neuter`), "la donna" (*"the woman"*, `gender:female`).

If the language doesn't have a grammatical gender for non-pronominal noun phrases, but noun phrases might be substituted with personal pronouns with a grammatical gender, then the `gender` takes the value of the gender of the personal pronoun. For instance, "the woman" might be substituted by the pronoun "she". Then the `gender` feature takes the value `gender:female`.

Coordinated NPs in which each of the coordinated noun phrases has a different grammatical gender, like in *"Brothers and sisters"*, *"your son and your car"*, *"la sedia e il tavolo"* (*"the chair and the table"*) the `gender` feature for the markable is annotated with the value `gender:underspecified`.

Another use of the value `gender:underspecified` is the annotation of noun phrases that can refer to male or female entities in the real world, lie *"the Professor"* or *"the visitors"*.

In the implementation of this proposal to languages in which grammatical gender does not exist, like in the case of Basque, the value `gender:underspecified` will we used as default for all the markables. If the language has other possible values for the `gender` feature, the set of values will be extended.

Number

The possible values for this feature are `singular`, `plural`, `mass` and `underspecified`.

The value `mass` is used to annotate uncountable nouns, like `information`.

The value `underspecified` is used for the annotation of cases of coordination in which the coordinated noun phrases have different values for the `number` feature, like in the case of *"Mary and her parents"*

Person

The possible values for this feature are `first`, `second`, `third` and `underspecified`.

The value `underspecified` is used for the cases of coordination of noun phrases of different persons, like *"you and me"*.

3.2.2 MIN_words and MIN_IDs

The annotation scheme introduces a MIN_ID attribute to help in the alignment of the hand annotated markables with markables produced by systems that use the corpus resources for their training and evaluation. The annotation of MIN_IDs was not part of the MATE proposal and was not present in ARRAU and the VENEX corpora.

If the head of the noun phrase is a common noun, then this noun is the MIN_word of the markable.

(21) `[The` **`head`** `of the nation 's largest car-dealers group]`

If the head of the noun phrase is a named entity, then the full named entity as a list of words is the MIN_word of the markable.

(22) **`[Robert Erwin`** `, president of Biosource]`

(23) **`[El Paso Refinery Limited Partnership`** `, El Paso , Texas , ( ELP )]`

If the noun phrase have more than a single head, like often happens in coordinated noun phrases, we don't annotate the head. The reason is that there are two possibilities. The first one is to annotate it using a similar approach than in the annotation of MUC, like in example (24). But the annotated list of words doesn't correspond to any linguistic category.

(24) `[the two` **Croatians and Brown**`]`

The second approach would be to annotate it as a list of heads, but then we would have a discontinuous constituent like we can see in example (25), what is not desired for the markable identification and alignment.

(25) `[the two` **Croatians** `and` **Brown**`]`

3.2.3 Annotation of language specific features

The present annotation proposal has been used for the annotation of English and Italian data. There are two specific phenomena of the Italian that require a specific treatment, the clitics incorporated to the verb and the phonologically empty subjects.

Incorporated clitics

The annotation scheme proposal has been implemented for English and Italian. In Italian we have two features that need a specific treatment: the clitic pronouns attached to the verb and the phonologically empty subjects.

An phenomenon to be considered in Italian and other romance languages is the presence of clitic pronouns morphologically incorporated to the verb like in example (26).

(26) ...[Il giudice] [gli] nego' [questa richiesta] e procedette invece ad acquistare [alcuni indumenti da [fargli] indossare]...[2]

In *fargli*, the clitic *gli* (*to him*) is attached to the verb *fare* (*make*).

This anaphoric expressions have two main differences to the previously discussed kinds of markables:

[2] *The judge to-him rejected this request and proceeded instead to buy some clothes to make-to-him wear.*

1. The syntactic category of the anaphora. In opposition to the other markables, in which the anaphoric expression was a nominal expression, in this case the anaphora is part of the verbal complex.

2. More than one anaphora can appear in the same token, as in example (27)

(27) ...[[dammelo]][3]

In *dammelo*, the clitics *me* (*to me*) and *lo* (*it*) are attached to the verb *dare* (*give*).

A possible way to handle with this phenomena is the use of a morphological decomposed representation of the verb produced by a linguistic tool, but this kind of resources are at the moment too limited to be reliably used for most languages.

Another way to solve the problem is the solution given by the MATE proposal presented in section 2.4, the introduction of a tag for markables that are not realized by nominal expressions. That includes verbs with incorporated clitics together with other kind of markables like the antecedent of discourse deixis.

I propose a solution similar to the proposal of MATE, but limited to the case in which the markable is a verb. To do it the scheme has to introduce at the begining of the annotation process a distinction between two types of markables, **nominal markables** for nominal expressions and **verbal markables** for cases in which the verb will be selected as a markable.

If the verb has more than one incorporated clitic pronoun, for each of them the annotator will create a separate markable in the annotation.

Empty subjects

In Italian the subject of a sentence can be empty as in example (28). From the point of view of the annotation this phenomenon is problematic since empty subjects are not at all realized in the surface form of the text, and there is no token than can be included in the markable.

[3] *give-me-it*

(28) ...Cesare Battisti ... e' stato un geografo, politico e irredentista italiano. ∅ Nacque in Trentino...[4]

In example (28) the subject of the second sentence is a morphologically null reference to the subject of the first sentence.

A possible way to annotate empty pronouns might be to add traces as tokens in the corpus with help of a dependency parser, what like in the previous case is not suitable for most languages in which empty pronouns have an anaphoric function.

The solution that I adapt here is to use the **verbal markables** proposed for the annotation of clitics (as in example (29), specifying with an extra tag that it is a case of empty subject.

(29) ...[Cesare Battisti] ... e' stato [un geografo, politico e irredentista italiano]. [Nacque] in [Trentino]...

This proposal adopted for Italian should be extended in order to be able to cover other languages, like Spanish or Portuguese, in which subject and object might be realized by empty pronouns, or more complex cases in other languages, in which different constituents can be elliptized.

3.3 Grammatical function

Studies about salience of antecedent candidates in anaphora resolution as (Brennan, Friedman, and Pollard, 1987) and (Lappin and Leass, 1994) show that in languages like English the grammatic role of the antecedent plays a relevant role in the resolution process.

Here I present a set of grammatical functions to use in the annotation of the corpora. Although roles like Subject or Object seems to cover all the languages, the modification of other roles like the genitive might be necessary to adapt the scheme to different languages.

1. **Subject:** The mention represented by the markable is the subject of a sentence.

2. **Object:** The mention represented by the markable is the direct object of a sentence.

[4] *Cesare Battisti... was an Italian geographer, politician and 'irredentista'. (He) was born in Trentino ...*

3. **There-object:** The mention represented by the markable is the object of a sentence like in example (30)

   (30) there are **three people** under the table".

4. **Complement:** The mention represented by the markable is the indirect object of a sentence.

5. **Adjunct:** Adjuncts and complements that are not direct or indirect object.

6. **predicate:** The mention represented by the markable has a predicate function, that is, expresses a property of an entity. These markables can be part of a copulative sentence or appear in an appositive construction.

7. **NP-complement:** The mention represented by the markable is a complement of a noun phrase. They are usually connected in English by the preposition *of*.

   (31) an average life of **eight years**

8. **NP-modifier:** The mention represented by the markable is a modifier of a noun phrase. They can appear in pre-modifier position like in example (32) and (33) or as prepositional phrase like in example (34)

   (32) The **Kuala Lumpur** stock exchange

   (33) the **Nigerian** president

   (34) the lady with **red hut**

9. **Genitive:** The mention represented by the markable is linked by a genitive like in (35) or it is a possessive adjective like in (36) and (37).

   (35) **Mary's** car

   (36) **her** car

   (37) la **sua** machina

10. **Adjective-modifier:** The mention represented by the markable modifies an adjective.

(38) equivalent in **value**

11. **NP-part:** The markable that is being annotated is embedded in another markable. The mention represented by the embedding markable is a part of the mention that is being annotated, as one can see in example (39)

    (39) Many of **the nation's highest-ranking executives**

3.4 Reference and no-reference

When the surface features of the markable are annotated, the next step consists of the distinction between referring and non referring noun phrases.

Referring noun phrases are phrases that to refer entities in the real world (Webber, 1979). The non referring noun phrases are classified in the following five types:

1. Expletives
2. Predicative
3. Quantified expressions
4. Coordinated noun phrases
5. Idiomatic expresions

3.4.1 Expletives

Expletives are a case of noun phrases that don't refer to real world entities.

(40) [**There**] are two people waiting for the interview.

(41) [**It**] is always nice to see you.

(42) [**C**]'è Mario al telefono

The same words can be used with an anaphoric sense, like in:

(43) What bugs me the most about this book is readers reactions to [**it**]. If the fact that [**it**] is a bestseller does nott cause me enough pain

(44) The new car is [**there**]

(45) Napoli è vicina. Vi [**ci**] porto.

In example (45) the pronoun *"ci"* is coreferring with *"Napoli" (Naples).*

(46) **Ci** piace andare al mare

In example (46) *"ci"* is the first person plural personal pronoun (*"we"*)

3.4.2 Predicates

Predicates are noun phrases that don't introduce a new entity in the discourse or corefer with previously mentioned entities. Their function is rather to express a property of other noun phrases.

Copulative sentences are a good example to show how predicates are not referring expressions. As one can see in example (47), if "a professor" would be a referring expression, it should be coreferring with the subject of the copulative sentence Mary, what is linguistically wrong.

(47) **Mary** is a professor.

Sometimes is difficult to decide which of the noun phrases connected by a copula is the referring expression. The proposed criteria are:

- If one of the noun phrases a named entity and the other isn't, the named entity is referring.

  (48) The new professor is **Dr. Mary Smith**.

- If one is a personal pronoun and the other isn't, the personal pronoun will be annotated as referring.

  (49) **She** is the new professor.

- If one of the noun phrases is a definite noun phrase and the other one isn't, the definite noun phrase will be annotated as referring.

  (50) A nice woman is **the new professor**.

- In other case the constituent in the subject position is annotated as referring in case of English and Italian. In the implementation of this scheme in other languages language specific criteria should be defined.

(51) **The lady that you meet yesterday** is the new professor.

Another construction in which noun phrases have often a predicative function are the appositions.

As explained in section 3.1 in a appositive construction, a noun phrase usually refers to an entity and the other noun phrase is a predication about it.

(52) Giorgio Napolitano, president of the Republic.
`[[`**Giorgio Napolitano**`], president of the Republic.]`

(53) The president of the Republic, Giorgio Napolitano.
`[The president of the Republic, [`**Giorgio Napolitano**`]].`

In example (52) and (53) the markable [Giorgio Napolitano] will be annotated as referring, and the complete noun phrase as predicate.

There are cases of apposition like in example (54) in which all the markables, appositive or not, are referring expressions.

(54) IBM, Pasadena, Calif.
`[IBM, [Pasadena, [Calif.]]]`

3.4.3 Quantifiers

Quantified noun phrases are used to indicate the proportion of elements of a set that have a concrete property, or the identity of these elements. The annotation scheme considers quantified noun phrases as non referring expressions.

A kind of quantified noun phrases used to determine the identity of the elements of a set that fill a condition are the wh-noun phrases like in example (55) and (56).

(55) and then [**which route**] do you want to take?

(56) [**where**] is the boxcar?

In these examples the speaker does not refer to any object or location in the real world.

Other kind of quantified noun phrases are the quantifiers with “all”, “all of”, “any” “every”, “each”, “how many” etc. are used to indicate the quan-

tity of elements of a set that share a property like in examples (57) and (58).

(57) [**How many oranges**]?

(58) [**Every TV network**] reported its profits yesterday.

Although the quantified NP is not anaphoric, the domain of quantification will be annotated as referring as one can see in:

(59) and then [which [**route**]] do you want to take?

(60) [How many [**oranges**]]?

(61) [Every [**TV network**]] reported its profits yesterday.

3.4.4 Coordination

Coordinated NPs are not referring expressions, but the coordinated items are.

(62) [[**US**] **and** [**Mexico**]] have signed new commercial agreements.

For instance in example (62) the markables [US] and [Mexico] are annotated as referring, and the coordination as non referring.

3.4.5 Idioms

Noun phrases that occur inside of idiomatic expressions don't refer to any object in the real world and don't contribute to the meaning of the expression. For instance in the expression of example (63) the meaning of the expression doesn't derive from the meaning of "the neck" or "the nape".

(63) by [the nape of [the neck]]

3.5 Semantic type

All the referring mentions including pronouns are annotated with information about semantic type. The annotation scheme is strongly inspired by the ACE guidelines for the Named Entity Recognition Task (LDC, 2004) and the annotation guidelines of the MUC-7 for the numeric expressions (Chinchor, 1997). In addition, we introduce a distinction between concrete and abstract entities, and a category of animated entities.

1. **Person:** The markable refers to a person or group of persons like families and coordinated markables (examples (65) and. This semantic type is restricted to the annotation of human beings. Personifications of animals like in (67), mechanic artifacts like in (68), etc. will be annotated with the tag Animate or Concrete.

   (64) [**The speaker**]

   (65) The tales of the [**Grimm brothers**]...

   (66) [[John] and [Mary]] live with [**their**] parents.

   (67) Then said [**the Wolf**] , so cunning, "What is it that you bear?"

   (68) "It is not right," said [**the robot**]. "We were made to serve all."

2. **Animate:** The markable refers to living beings that are not persons.

3. **Organization:** Formally constituted organization, like companies, public institutions, sport teams, political, cultural and religious organizations, etc.

4. **Facility:** The markables are human-made structures like buildings, bridges, factories, etc. Markables referring to the transportation structure, like tunnels, bridges, railways, stations, roads or airports will be annotated with this category.

5. **Geopolitical entity:** The markable refers to politically defined geographical areas, like countries, cities or regions. This category merges the geographical region, the ruling institution and the inhabitants as recommended in the ACE guidelines. The main reason to decide for the introduction of this category is the low reliability in the distinction between the cases in which the markable of the GPE refers to a function as location or as organization.

6. **Location:** The markable refers to places that are not geopolitical entities. This category covers geographical entities like mountains, rivers, seas, etc, streets, or postal addresses. It is used too to annotate web sites, e-mail addresses and telephone numbers.

7. **Temporal:** It is used to annotate expressions that refers to time points and intervals, like dates, years, or time units like "a week".

8. **Numerical:** It is used to annotate percentages and prices. This category merges the `percent` and `money` categories of the annotation guidelines for numeric expressions (`NUMEX`) of the MUC-7.

9. **Concrete:** The markable refers to physical inanimate entities that don't belong to one of the previously mentioned categories.

10. **Abstract:** Used to annotate expressions that refers to states and other abstract entities that are not events or temporal expressions, e.g. the justice, the law, philosophy, etc.

11. **Plan:** Used to annotate expressions that refer to plans, eventualities and nominalizations.

12. **Other:** If the semantic type does not belong to this list.

13. **Unknown:** If the annotator is not able to identify the semantic type.

3.6 Information status

The markables are classified in two categories, `new` if the markable is the first mention of the entity, or `old` it the entity has been introduced previously in the discourse.

- **New:** The markable refers to the first mention of an entity in the discourse.

- **Old:** The markable refers to an entity or abstract object that have been previously mentioned in the discourse.

3.7 Anaphoric links

In this step the annotator has to annotate the links between the markables that have been marked with the value `old` for the tag information status, and their antecedents.

There are two kinds of antecedent:

- **Phrase:** when the antecedent refers to an object that has been mentioned using a markable.

(69) [**Commonwealth Edison Co.**] was ordered to refund about [$ 250 million] to [[**its**] current and former ratepayers] for [illegal rates collected for [cost overruns on [a nuclear power plant]]]

- **Segment:** This value is used if the markable refers to abstract objects like events, actions or facts that have been discussed in the discourse, but not referred using a markable.

(70) Recently , [the boards of [both [the parent company] and [the thrift]]] also **voted to suspend [dividends on [preferred shares of [both companies]]] and convert [all preferred] into [common shares]** .
[The company] said [**the move**] was necessary to meet [capital requirements] .

3.7.1 Annotation

The annotation of anaphoric reference consists on a link from the markable of the anaphora to the antecedent.

The fist step of the annotation process consists on distinguishing whether the antecedent is another nominal expression as in example (69) or a segment of the discourse like in example (70)

Discourse deixis

If the anaphoric expresion refers to a segment of the previous discourse, the annotation consists in link the markable of the anaphora to the sentences that the anaphora refers. For instance, in example (70) the full sentence *"Recently, the boards of both the parent company and the thrift also voted to suspend dividends on preferred shares of both companies and convert all preferred] into common shares"* should be marked as antecedent of the anaphoric expresion *"the move"*.

Sentences have not been selected previously as markables for the annotation, then the annotation scheme should implement a mechanism for the markup of the span of the antecedent. In the next chapter I describe an implementation of an extra level used for the annotation of discourse deixis.

Multiple antecedents

A markable can refer to a single markable or to a set of markable, that is what we call plural markables. For instance in example (71) the markable [The partners] refers to both business partners, but not to the coordination.

(71) [**Mrs. Park**] and [**Mr. Kim**] opened a new business. [**The partners**] have invested $1,000,000 in the joint venture

The plural markable is annotated with the tag `phrase_antecedent:multiple_phrases` and with a link to each of the antecedents.

Markables with a single antecedent will be marked with `phrase_antecedent:single_phrase`.

Ambiguity

As (Poesio and Artstein, 2005) points not always is only a unique interpretation for a markable, like occurs with the interpretation of the pronoun it in examples (72) and (73).

(72) Be careful hooking up [**the engine**] to [**the boxcar**] because [**it**] is faulty

(73) [**The house**] is on [**a long street**]. [**It**] is very dirty.

In example (72) the pronoun *"it"* might potentially corefer with *"the engine"* and *"the boxcar"*. In example (73) the situation is similar. In a possible interpretation the pronoun **"it"** would corefer with *"the house"*. This interpretation competes with a second one in which *"it"* corefers with *"a long street"*.

If a markable has alternative interpretations it will be annotated with the tag `ambiguity:ambiguous`, and annotated with separate links to each interpretation.

The given examples show ambiguity between two possible antecedents for a markable that have been previously tagged with information status `old`. But another possibility is that a interpretation is that the markable corefers with a previous markable and the alternative interpretation is that the markable is the first mention of an entity (`new`) or a non referring expression.

If there is a unique interpretation of the markable, then it will be marked wit the tag `ambiguity:unambiguous`.

3.7.2 Annotation of related object

The attribute `related_object` is used to mark anaphoric relations between objects different than coreference like the bridging relations described in (Passonneau, 1997).

As mentioned in (Poesio, 2004b) bridging relations are difficult to be detected and reliably annotated.

We have constrained the annotation to three relations:

- Set membership: the markable refers to an object which is one element in the set of objects that the antecedent refers to.

  (74) Giorgio ha due fratelli. **Il più grande** va all'Università

- Part-of : the markable refers to a part of the object the antecedent refers to.

  (75) Quella mattina, Mario prese la macchina perché era in ritardo, ma lungo la strada bucò **le ruote**.

- Attribute: the markable refers to an attribute of the object which the antecedent refers to.

  (76) Finalmente Luca si è comprato una macchina nuova. **Il colore** non è dei più belli.

3.8 Conclusions

In this section I have provided a framework proposal for the implementation of annotation schemes to annotate anaphoric information in text.

I have presented criteria for the identification of markables, and to distinguish which of the markables have a referring function and are candidates to be anaphora or antecedent, and which are not.

I have discuss the determination of the markable boundaries, introducing the concept of discontinuous markables, a concept that was not used in the previously presented corpora.

A set of general and language specific features to add information to the markables, like morphosyntactic agreement features, semantic features, grammatical role or mechanisms to do possible the annotation of empty pronouns and clitic pronouns attached to the verb.

Finally I give criteria for the annotation of different types of anaphoric links, like identity relations, bridging relations and discourse deixis.

The implementation of this scheme for the annotation of English and Italian data and the produced corpora are described in the next section.

Chapter 4

Implementation of the proposal for the annotation of Italian and English data

In this chapter I describe how the proposal for the annotation of corpora presented in the previous chapter has been implemented for the annotation of an English corpus, ARRAU (Poesio and Artstein, 2008) and the Italian Live Memories Corpus – henceforth LMC (Rodriguez et al., 2010).

In section 4.1 I discuss the methods to extract automatically markables for the human annotation directly from treebanks, like in the case of the English data, or from web pages in the case of the Italian corpus.

Section 4.2 presents MMAX (Müller and Strube, 2006), the annotation tool used for the human annotation of the text and its data format.

Section 4.3 describes how the annotation proposal presented in previous chapter have been implemented for the annotation of English and Italian data.

After that section 4.4 gives information about the annotation process and the training of the human annotators.

Section 4.5 offers a description of the annotated datasets and a reliability study of the annotation.

4.1 Extraction of text and markables

The starting point of the annotation is the automatic extraction of markables from the syntactic representation of the sentences of the texts. We have used different procedures for this first extraction of markables in the different English and Italian sets, and for the conversion of the text and extracted markables in the standoff format used by the annotation tool MMAX2.

4.1.1 Extraction of markables for the English data

One of the datasets are articles of the Wall Street Journal annotated with syntactic information in the Penn Tree Bank (Marcus, Marcinkiewicz, and Santorini, 1993). The syntactic trees have been used to extract the list of tokens and sentences and all the noun phrases of the texts.

To build the other datasets of ARRAU the procedure starts from the raw text, that have been tokenized and syntactically processed using the Charniak parser (Charniak, 2000), a maximal entropy based constituency parser that produces a format similar to the used in the Penn Treebank. Like in the case of the WSJ dataset, the syntactic trees have been used to extract the list of tokens, sentence boundaries and all the noun phrases.

4.1.2 Extraction of the Italian data

Extraction of the texts

The texts so far annotated for the LMC are taken from the Web. Web pages offer information in a visually structured way, structure that we keep as a part of the corpus. The preservation of layout information is especially relevant in the annotation of blog sites.

Wikipedia pages offer structured text, with titles and subtitles, structure that we aim to capture in our corpus.

Blog pages contain text introduced by the author and comments introduced by users of the blog. Comments reflect opinions about the main post, about other comments, and update of information. They are written by different authors with different writing styles. In the comments we find mentions to entities that corefer with entities appearing in the main post, or in other comments. Another interesting feature is the mention of other comments using IDs or the name of the user.

When using content from the web, subsequent usage might suffer from messy data – in our case, open or disguised advertising, and boilerplate are the main concerns. Consequently, the data need to be cleaned first. To this end, we use KrdWrd (Steger and Stemle, 2009), a tool for the unified processing of web content. Here, we can use it to select the parts of a web page we want to keep and also add annotation categories in a consistent way.

Extraction of markables

The extracted text is processed in a developed pipeline. At the beginning we use the tokenizer, part of speech tagger and sentence splitter from the toolkit TextPro (Pianta, Girardi, and Zanoli, 2008), which produces a tabular format.

Then we parse the data using the MALT dependency parser (Nivre et al., 2007) trained on a Italian treebank, the TUT (Bosco et al., 2000). Afterwards, we use the produced dependency trees to create markables for all noun phrases.

Finally we use the produced dependency trees to create markables for all noun phrases of the texts, and we export them and the list of tokens in the MMAX02 format.

4.2 The annotation tool

For the annotation of texts we use the MMAX annotation tool (Müller and Strube, 2006).

One of the main reasons to choose this tool is that it allows to annotate different kind of links, and visualize this different links in different colors, making easier to distinguish the different kind of relations that we annotated.

A further reason is the used standoff format. It allows to separate in different levels different kinds of information, like the boundaries of the sentences, the discourse units, the markables for the manual annotation. That makes possible to add new levels in the annotation, or to introduce changes in the existing levels, without to change the other levels or the basic data, what is represented as a list of tokens.

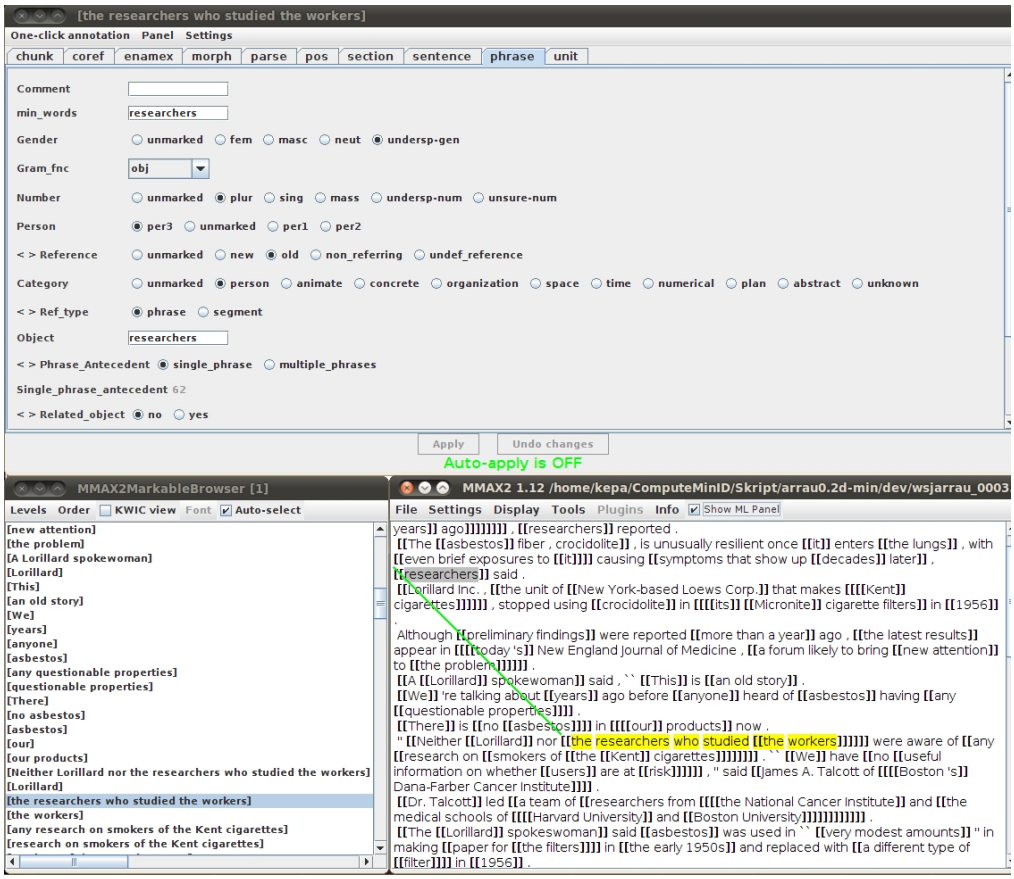

Figure 4.1: Annotation interface of MMAX

A last reason is the user friendly graphic interface of the tool. It makes easy the visualization and selection of markables in a markable browser, the ordering of markables in alphabetic order or in linear order, and the visualization and selection of the values for the attributes of the annotation in an attribute window.

4.3 Implementation of the scheme

The annotation scheme has been implemented as an XML file in the format of MMAX as presented in figure 4.3.

In this format the data is stored as a list of tokens in a XML file as in example (1).

(1)
```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE words SYSTEM "words.dtd">
<words>
<word id="word_1">Solo</word>
<word id="word_2">woodwind</word>
<word id="word_3">players</word>
<word id="word_4">have</word>
<word id="word_5">to</word>
<word id="word_6">be</word>
<word id="word_7">creative</word>
<word id="word_8">if</word>
<word id="word_9">they</word>
<word id="word_10">want</word>
<word id="word_11">to</word>
<word id="word_12">work</word>
<word id="word_13">a</word>
<word id="word_14">lot</word>
<word id="word_15">,</word>
<word id="word_16">because</word>
...
</words>
```

The different levels of the annotation are defined as lists of markables and stored in separated XML files. An example of annotation level is the sentence level, which contains the sentence splitting of the text, and we can see it in example (2).

```
<?xml version="1.0" encoding="UTF-8"?>
<annotationscheme>

<!-- introduction of min_words -->
<attribute id="min_words" name="min_words" type="freetext">
<value id="mins_word" name="Mins"/>
</attribute>

<!-- annotation of gender -->
<attribute id="gender" name="Gender" type="nominal_button">
<value id="no_mark" name="unmarked"/>
<value id="fem" name="fem"/>
<value id="masc" name="masc"/>
<value id="neut" name="neut"/>
<value id="undersp-gen" name="undersp-gen"/>
</attribute>
....
<!-- Annotation of reference and information status -->
<!-- open new options for each category -->
<attribute id="reference" name="Reference"
type="nominal_button">
<value id="no_mark" name="unmarked"/>
<value id="new_obj" name="new" next="category,object,
related_object,generic"/>
<value id="old_obj" name="old" next="category,ref_type,
related_object,ambig_choice,generic"/>
<value id="non_ref" name="non_referring" next="non_ref_type"/>
<value id="undef_ref" name="undef_reference"/>
</attribute>

<!-- Annotation of no referring markables -->
<attribute id="non_ref_type" name="non_ref_type"
type="nominal_button">
<value id="unknown" name="unknown"/>
<value id="expletive" name="expletive"/>
<value id="predicate" name="predicate"/>
<value id="quantifier" name="quantifier"/>
<value id="coordination" name="coordination"/>
<value id="idiom" name="idiom"/>
<value id="incomplete" name="incomplete"/>
</attribute>
...

</annotationscheme>
```

Figure 4.2: Fragment of the implementation of the annotation scheme for MMAX

(2)
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/sentence">
<markable id="markable_0" span="word_1..word_24" orderid="0"
mmax_level="sentence" />
<markable id="markable_1" span="word_25..word_68" orderid="1"
mmax_level="sentence" />
<markable id="markable_2" span="word_69..word_79" orderid="2"
mmax_level="sentence" />
<markable id="markable_3" span="word_80..word_121" orderid="3"
mmax_level="sentence" />
<markable id="markable_4" span="word_122..word_180" orderid="4"
mmax_level="sentence" />
<markable id="markable_5" span="word_181..word_203" orderid="5"
mmax_level="sentence" />
<markable id="markable_6" span="word_204..word_221" orderid="6"
mmax_level="sentence" />
...
</markables>
```

4.3.1 Differences between English and Italian annotation schemes

There are some differences in the schemes for English and Italian. The reasons are:

- The introduction of verbal markables to facilitate the annotation of attached clitic pronouns and empty subjects.
- We tried to use as much as possible of the previous annotation of the ARRAU files. The consequence is that there is some differences in the annotation of semantic types.

Introduction of verbal markables

The main difference between both corpora is that for Italian we introduce at the beginning the distinction between nominal and verbal markables. In the English annotation all the markables are nominal, and the annotator doesn't need to choose between these options.

If the annotator introduces averbal markable, then she/he has to choose between two possible values: clitic pronoun like in example (3) or empty

subject. After the annotation of the type of verbal markable the annotation proceeds in the same way for both corpora.

(3)
```
<markable id="markable_174" span="word_312" gender="masc"
related_object="no" phrase_antecedent="single_antecedent"
number="sing" category="facility" verbal_type="clitic"
ambiguita="non_ambiguo" reference="old" type_reference="phrase"
mmax_level="phrase" single_antecedent="markable_93"
markable_type="verbal" person="per3" min_ids="word_312"
min_words="rendendolo" />
```

Semantic types

In the list of values for semantic types, the annotation of ARRAU doesn't have the type GPE, category that have been used in the LMC. On the other side we realized that annotators had problems to distinguish between eventualities and abstract entities, being the annotation not very reliable. Due to this reason we merged the category of abstract and event in the Italian corpus.

Grammatical function

In the actual phase of the annotation the Italian data has not been annotated with semantic role of the markables. To compensate it we have defined an additional annotation level, the dependency relation (DEPREL) level.

The DEPREL level contains automatically predicted grammatical role for each token of the annotation. The necessary information to produce this level is extracted from the output of the Malt dependency parser.

(4)
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/deprel">
<markable id="markable_1650" span="word_1" mmax_level="deprel"
tag="ROOT"/>
<markable id="markable_1651" span="word_2" mmax_level="deprel"
tag="RMOD"/>
<markable id="markable_1652" span="word_3" mmax_level="deprel"
tag="PN"/>
<markable id="markable_1653" span="word_4" mmax_level="deprel"
tag="DET"/>
```

```
<markable id="markable_1654" span="word_5" mmax_level="deprel"
tag="CONTIN-DENOM"/>
<markable id="markable_1655" span="word_6" mmax_level="deprel"
tag="RMOD"/>
<markable id="markable_1656" span="word_7" mmax_level="deprel"
tag="PN"/>
<markable id="markable_1657" span="word_8" mmax_level="deprel"
tag="ROOT"/>
<markable id="markable_1658" span="word_9" mmax_level="deprel"
tag="DET"/>
<markable id="markable_1659" span="word_10" mmax_level="deprel"
tag="ROOT"/>
<markable id="markable_1660" span="word_11" mmax_level="deprel"
tag="RMOD"/>
<markable id="markable_1661" span="word_12" mmax_level="deprel"
tag="PN"/>
<markable id="markable_1662" span="word_13" mmax_level="deprel"
tag="OPEN-PARENTHETICAL"/>
<markable id="markable_1663" span="word_14" mmax_level="deprel"
tag="RELCL"/>
<markable id="markable_1664" span="word_15" mmax_level="deprel"
tag="RMOD"/>
...
</markables>
```

4.4 Annotation process

The annotation of the corpus has two phases that required separated training:

1. The identification and correction of the correct boundaries of the markable.

2. The annotation of the markables with the set of features and the anaphoric links.

4.4.1 Correction of markable boundaries

The correction of the automatically determined markable boundaries is one of the most difficult and time consuming tasks of the annotation, and especially in the case of coordination and discontinuous markables requires non

only syntactic, but some time semantic reasoning from side of the annotator.

The main issues that human annotators have to do are:

- Correction of boundaries if a noun phrase contains a post-modifier. Markables extracted from the syntactic annotation of the Penn Treebank have the following structure, that follows the syntactic annotation of the corpus.

  (5) [[the man]from [London]]

  Here the annotator have to remove the markable of the head of the post-modified noun phrase, producing.

  (6) [the man from [London]]

- Correction of markable boundaries in coordinated noun phrases. The output of the parser doesn't produce the discontinuous markables discussed in previous chapter, and the annotator have to introduce them by hand. That is one of the most time consuming part of the annotation.

- Introduction of markables for non restrictive relative phrases The parser is not able to distinguish between relative pronouns at the beginning of restrictive and non restrictive relative phrases, and the annotator have to introduce them by hand.

4.4.2 Agreement features and grammatic role

When the annotator begin with the manual annotation, first of all he/she has to annotate the morphosyntactic agreement features `Gender`, `Number` and `Person`.

The annotation of this group of features is usually not problematic for most annotators with the exception of the use of the tag `underspecified` for the feature `Gender` in English. At the beginning most annotators tend to confuse `underspecified` with `neuter`.

4.4.3 Reference and information status

When the morphosyntactic agreement features have been annotated the annotator has to decide whether a markable is referential or not, and if it is

referential, whether it is the first mention of the markable or it has been previously mentioned.

In order to speed up the annotation I have grouped both steps in one. In this way the annotator has to tag the markable with the tags `old`, `new` and `non_referring`.

This feature is usualy easy to be annotated, but sometimes there are some disagreement to determine whether the correct tag is `new` or `non_referring`. The distinction between `new` and `old` is usualy no problem for the annotators.

Another observation is that the efficient distinction between `old` and `new` requires that the annotator should be able to annotate the text in a session, or at last in one day. Annotators that worked in long texts, most of them of the Italian Wikipedia, reported that it was difficult to keep in mind whether entities that were not often mentioned in the text were old or new. That happened more when the markables of anaphora and antecedent were not named entities, and the entities were realized with different strings.

4.4.4 Semantic type

The annotation of this feature were relatively fast for no pronominal noun phrases. In case of pronouns the required interpretation made the annotation more difficult.

In the annotation of the Italian corpora we detected cases of confusion between the annotation of `gpe` and `org`, and between `concrete` and `abstract`.

In the annotation of ARRAU the main source of disagreement war the distinction between `abstract` and `event`.

4.4.5 Type of antecedent and annotation of anaphoric links

If a markable is marked with the tag `old` the annotator has to tag the markable with the type of antecedent. There are two kinds of antecedents: `phrase_antecedent` if the antecedent is another markable, or `segment_antecedent` if the antecedent is a segment, as in example (7).

(7)
```
<markable id="markable_128" span="word_228..word_229"
generic="generic-no" person="per3" related_object="no"
```

ambiguity="unambiguous" gram_fnc="subj" number="plur"
type="unmarked" reference="old"
segment_antecedent="unit:markable_343;unit:markable_338"
category="abstract" mmax_level="phrase" **ref_type="segment"**
gender="neut" min_words="remarks" min_ids="word_229" />

Link to phrase antecedent

If the markable has been tagged with `phrase_antecedent` the annotator makes a pointer to the antecedent using the MMAX user interface as showed in figure 5.1.

The anaphora can have an only antecedent as in the example presented in figure 5.1, or more than one antecedent as in example (8) and figure 4.3.

If the markable has a single antecedent, the annotator will annotate the attribute `phrase_antecedent` with the value `single_phrase_antedecent` and make a link from the markable of the anaphora to the markable of the antecedent using the MMAX annotation window as showed in figure 4.3.

If the markable has multiple antecedents, the annotator will annotate the attribute `phrase_antecedent` with the value `multiple_phrase_antedecent` and make a link from the markable of the anaphora to all markables of the antecedents. For instance in example (71) the anaphora *"the two"* has as antecedent the nominal expressions *"California Plant Protection* and *"Pinkerton"*.

(8) Yet although **California Plant Protection** was netting bigger and bigger clients the firm provided security for the 1984 Summer Olympics in Los Angeles it still did n't have the name recognition of Pinkerton 's.
...
He decided he could easily merge **Pinkerton** 's operations with his own while slashing overhead costs because **the two** already operated in many of the same cities.
<markable id="markable_23" span="word_359..word_360"
generic="generic-no" person="per3" related_object="no"
ambiguity="unambiguous" gram_fnc="subj" number="plur"
reference="old" **phrase_antecedent="multiple_phrases"**
category="organization" mmax_level="phrase"
multiple_phrase_antecedents="markable_162;markable_140"

ref_type=”phrase” gender="neut" min_words="two"
min_ids="word_360" />

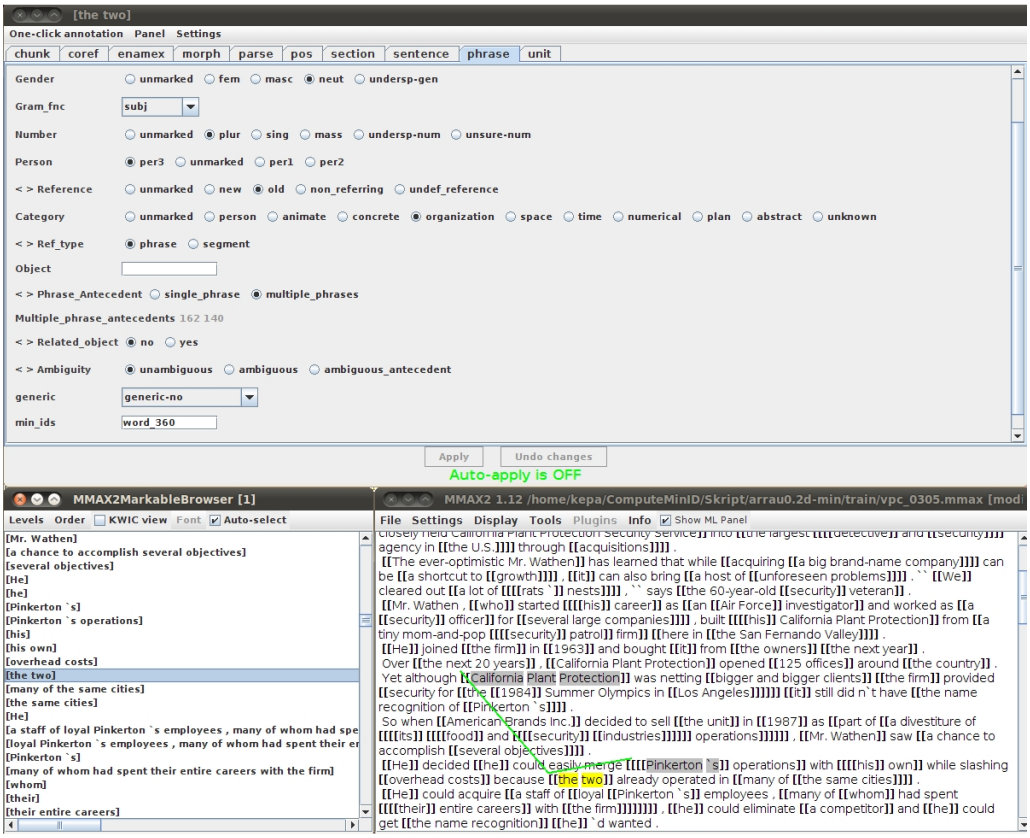

Figure 4.3: Annotation of multiple antecedent

Link to segment antecedent

If the markable have been annotated with `segment_antecedent` then the annotator has to select the segments of the discourse that serve the antecedent to the anaphora. To segment the discourse in these possible antecedents we use two different parser utilities, the Berkeley parser for English and the MALT dependency parser for Italian. Then the segments that have the node `S` in the syntactic tree will be considered as segments of the discourse and stored in an extra level, the UNIT level presented in example (9).

(9)
```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
```

```
<markables xmlns="www.eml.org/NameSpaces/unit">
<markable id="markable_352" span="word_278..word_295"
mmax_level="unit" subject="unmarked" verbed="unmarked"
utype="unmarked" finite="unmarked" />
<markable id="markable_353" span="word_280..word_294"
mmax_level="unit" subject="unmarked" verbed="unmarked"
utype="unmarked" finite="unmarked" />
<markable id="markable_354" span="word_285..word_294"
mmax_level="unit" subject="unmarked" verbed="unmarked"
utype="unmarked" finite="unmarked" />
<markable id="markable_391" span="word_647..word_657"
mmax_level="unit" subject="unmarked" verbed="unmarked"
utype="unmarked" finite="unmarked" />
...
</markables>
```

The annotator has to select the markables of the UNIT level that serve as antecedent of the anaphora as showed in figure 4.4.

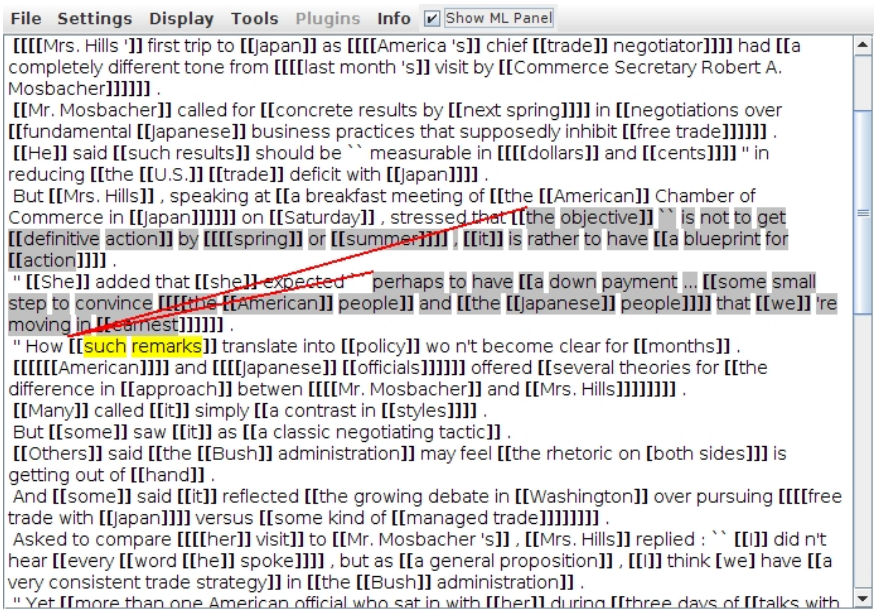

Figure 4.4: Annotation of discourse deixis in MMAX

Annotation of bridging references

If the markable is related to a previous markable by a relation other than identity, the annotator markes the markable with the value `yes` for the attribute `related_object` as in example (10).

Then the annotator do a pointer to the markable that represents the related object, and finaly gives a type to the link.

(10) The vast majority of **the U.S. corn** crop now is grown from hybrid seeds produced by seed companies.
A similar technique is almost impossible to apply to **other crops, such as cotton, soybeans and rice**.[1]
<markable id="markable_136" span="word_294..word_303"
generic="generic-no" person="per3" **related_object="yes"**
related_rel="other" gram_fnc="adjunct" number="plur"
reference="new" category="concrete" mmax_level="phrase"
related_phrase="markable_27" gender="neut" min_words="crops"
related_rel="other" min_ids="word_295" />

4.4.6 Ambiguity

In case of ambiguity the annotator has to tag the markable with the tag `ambigous`. That opens the possibility in the interface of annotating a second interpretation of the markable as showed in figure 4.5, that shows the annotation interface of MMAX in the annotation of example (11).

In the annotation of example (11) the annotator found two plausible interpretations for the markable *"his staff"*. The first interpretation is that the value of the information status for the markable is `old` and it corefers with the markable *"the in-house litigators*. The second possible interpretation is that the value of the information status for the markable is `new` and that *"the in-house litigators"* refers to a subset of *"his staff"*.

(11) Among the types of cases **the in-house litigators** handle are disputes involving companies doing business with GM and product-related actions, including one in which a driver is suing GM for damages resulting from an accident.
Mr. Pearce has also encouraged **his staff** to work more closely with GM 's technical staffs to help prevent future litigation.[2]

[1] vpc_0209
[2] wsjarrau_0617

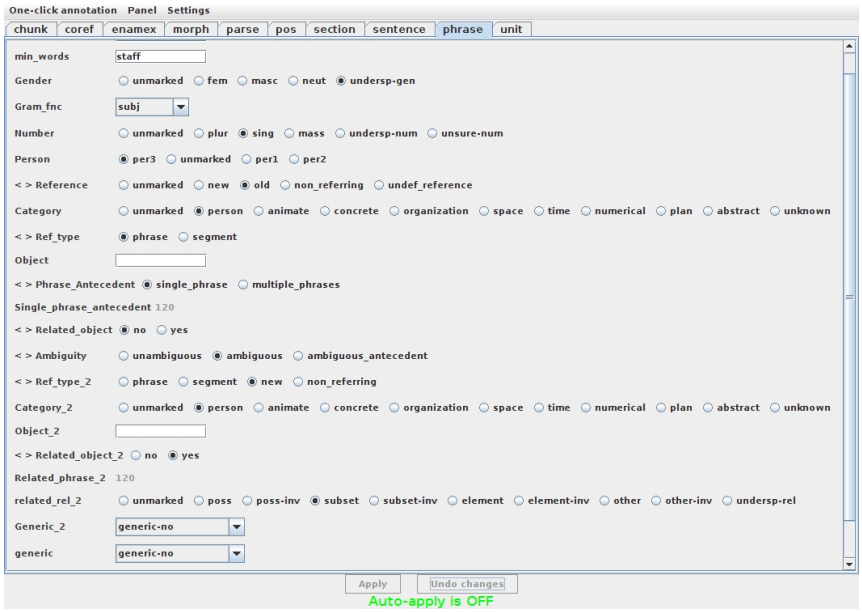

Figure 4.5: Distribution of the non-referring expressions in the ARRAU WSJ dataset

```
<markable id="markable_140" span="word_193..word_194"
generic="generic-no" generic_2="generic-no"
person="per3" related_rel_2="subset-inv"
related_object="no" category_2="person"
ambiguity="ambiguous" gram_fnc="subj" number="sing"
reference="old" phrase_antecedent="single_phrase"
related_object_2="yes" category="person"
mmax_level="phrase" related_phrase_2="markable_120"
ref_type="phrase" gender="undersp-gen"
comment="ambiguity:  whether whole set or subset"
single_phrase_antecedent="markable_120"
ref_type_2="new" min_words="staff"
min_ids="word_194" />
```

4.4.7 Main difficulties reported in the annotation

We don't have formal studies about the difficulty of the phases of the annotation, but most annotators agreed that the most difficult and time consuming task consisted in the correction of markable boundaries, task that sometimes take more time than the annotation of the markables with the set of features and the anaphoric links. That emphasizes the necessity of focus the research

on more elaborated markable extraction techniques that allow to reduce the effort of the human annotators.

An additional problem more related to a concrete dataset is the technical business language often used in the Wall Street Journal, language that often is unknown for annotators and designers of the annotation scheme. The problems to understand the meaning of these pieces of text has often two consequences:

- Problem to assign a correct semantic type.
- Incorrect detection of anaphoric links related to the previous issue.

To resolve it we used different sources, like Wikipedia entries and online sources about economy.

Another difficult task consisted of the detection of bridging relations. As reported by other annotation studies as (Poesio and Vieira, 1998), the agreement between annotators is usualy very low – $K = 0.24$ in the mentioned study. We have observed that different annotators tend to find different cases of bridging in the same text, most of them correct.

Training of the annotators

Annotators are individually trained and the training has the following phases.

- **Presential training**. In this phase trainer and annotator work together on the annotation. The annotator reads and discusses the annotation instructions with the trainer. Then the annotator has to correct the boundaries of the markables under supervision of the trainer and discussing the decisions made for each markable. It takes between 60 and 90 minutes.
- **Strongly supervised annotation**. The annotator get a set of files with around 4000 words and she/he has to correct the markable boundaries, introduce markables that have been not detected, etc. This annotation is strongly supervised, and feedback sent to the annotator.
- **Meeting** with the annotator to discuss the annotation mistakes and correct them.
- **Autonomous annotation work**.The annotator get a list of files and send each corrected files to the trainer. The trainer supervised them and send feedback, that should be used to correct the annotation.

When the annotator is confident with the instructions for the markable identification and correction in the markable boundaries, she/he becomes a new training to learn how to annotate the features and the anaphoric links in a training with similar structure to the previous one.

4.5 Description of the annotated data

4.5.1 The ARRAU corpus

The ARRAU corpus is a corpus of English text and dialogue. It follows closely the MATE annotation proposal.

The corpus includes the following datasets:

- Narrative stories from the Pear Stories Corpus (Chafe, 1980)
- Task oriented dialogues from the Trains-91 ((Gross, Allen, and Traum, 1993)) and Trains-93 ((Heeman and Allen, 1995)) corpora.
- News paper texts from the Penn Treebank WSJ dataset (Marcus, Marcinkiewicz, and Santorini, 1993).
- Texts of the Gnome Corpus (Poesio, 2004a)

The different datasets of the corpus were pre-annotated using different annotation schemes corresponding to different phases of the project. The annotation has been unified and optimized using the proposal of chapter 3.

Actually new data from the Penn Treebank is being added.

4.5.2 The LMC corpus

The Live Memories Corpus (henceforth LMC) is being developing in frame of the Live Memories project[3], a project that aims to collect, manage and integrate multimedia collective memories coming from different sources, and to scale up content extraction techniques.

The corpus consists of three genres.

- Wikipedia sites from the Italian Wikipedia[4].

[3]http://www.livememories.org
[4]`http://it.wikipedia.com`

- Blog entries with user comments.
- News paper articles from the regional news paper l'Adige[5].

The Wikipedia dataset of the LMC has been used for the multilingual coreference resolution task in the SemEval 2010 competition[6] (Recasens et al., 2010).

4.5.3 Statistics of the corpora

For the experiments reported in next chapter I have used two of the annotated sets, the WSJ dataset of ARRAU for English and the Wikipedia dataset of LMC for Italian.

The English data

The current set of annotated data of the WSJ dataset of ARRAU consists of 205 files with 147.6 K of tokens in 5585 sentences. In this set 47.9 K of markables have been selected for the annotation. Of the total of markables only a 1% were discontinuous markables.

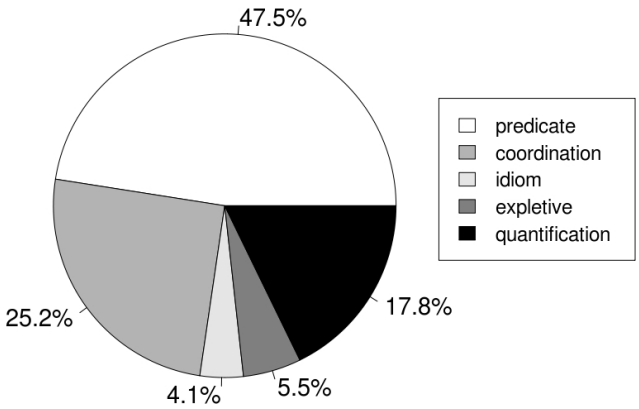

Figure 4.6: Distribution of the non-referring expressions in the ARRAU WSJ dataset

Figure 4.6 presents the distribution of non-referring markables in the dataset. 49.5% of the markables have been tagged as `discourse-new`, a

[5] `http://www.ladige.it`
[6] `http://stel.ub.edu/semeval2010-coref`

34.3% have been classified as `discourse-old` and a 12.6% as `non-referring` expressions.

The distribution of semantic types of the referring markables (figure 4.7) shows a predominance of references to abstract entities (35%), persons (17.5%) and organizations (16%).

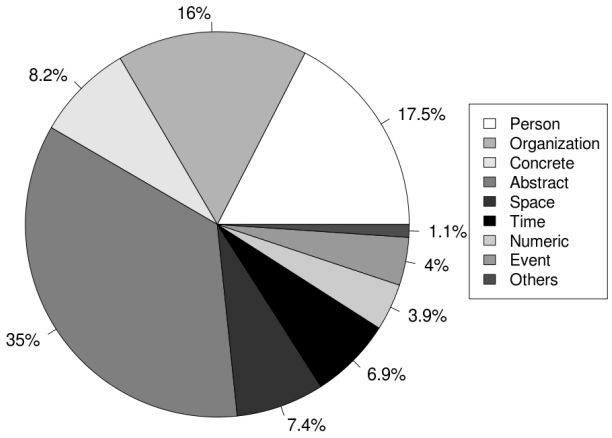

Figure 4.7: Distribution of semantic types of referring expressions in ARRAU WSJ dataset

The Italian data

The current set of annotated texts of Wikipedia consists of 144 files, with 140 K of tokens in 4703 sentences. In this dataset we have selected 44.5 K of markables for the annotation.

Of the total set of markables, 0.5% are discontinuous markables, and the same quantity (0.5%) are clitics attached to the verb. 4.5% of the anaphoric expressions are empty subjects, and all of them are linked to a previous antecedent. The resolution of this kind of anaphora is an interesting topic because they are part of the coreference chains.

57.8% of the markables have been tagged as `discourse-new`, 28.5% of the markables as `discourse-old` and 13.7% of the markables have been classified as `non-referring` expressions.

More than 50% of the `non-referring` markables of this dataset have been tagged as cases of predication, and 34% as cases of coordination. The distribution of categories can be seen on Fig. 4.8.

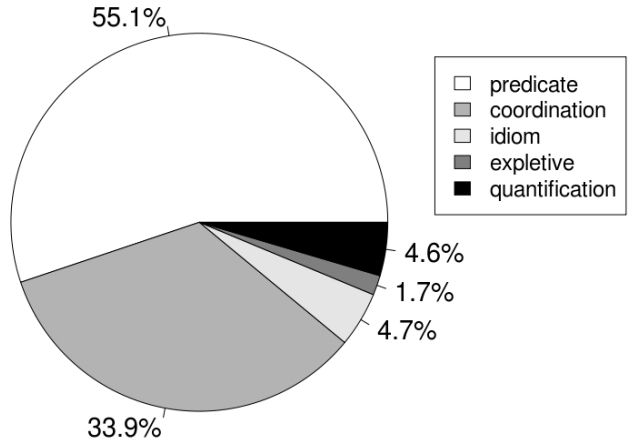

Figure 4.8: Distribution of the non-referring expressions in the Wikipedia dataset

In figure 4.9 one can see that the most common entity types are abstract (30.4%) and person (22.8%).

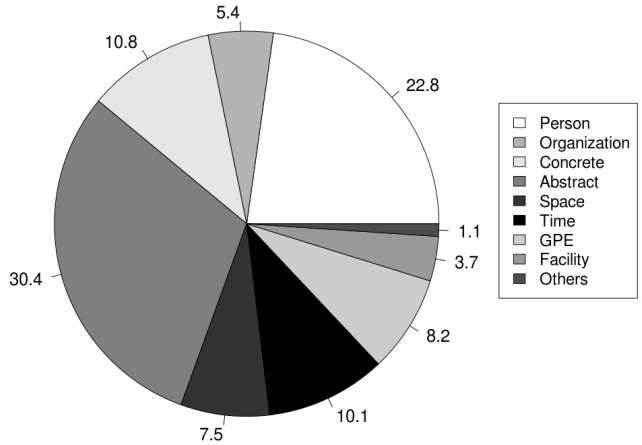

Figure 4.9: Distribution of semantic types of referring expressions in Wikipedia dataset

The main difference in the distribution of semantic types in both corpora is the higher precentage of entities of type organization in ARRAU (16%,

5.4% in LMC) and the higher precentage of entities of temporal entities in LMC (10.1%, 6.5% in ARRAU).

The differences correspond to the specificity of the different domains of the corpora. The texts of ARRAU are mostly related to business, domain in which entities of type organization referring to companies and institutions are usual.

In the Wikipedia domain we have selected mostly texts related to places, persons and historical items of the Trentino-Südtirol. The presence of historical items and biographies explain the the higher use of temporal entities.

4.5.4 Reliability studies

Annotation of ARRAU

We have not performed reliability studies for the annotation of the ARRAU corpus, then I report the results published in (Poesio and Artstein, 2008).

In the reported reliability experiments 20 annotators worked independently on the same text, and then computed the reliability measure α (Krippendorff, 80).

The results for the coreference chains were in the range of α between 0.6 and 0.7, values that can be considered as acceptable for the task. The α value for the annotation of discourse deixis was 0.55, reflecting the higher difficulty of it.

Annotation of the LMC

We have tested the reliability of our annotation scheme carrying out separate agreement studies for several features of the annotation scheme. For these studies we have used the Kappa coefficient (Carletta, 1996) between two annotators.

The first set of studies measures the **reliability of the annotation of features on the markable level**. These features are the information status, the referentiality of the markable and the semantic type of referring expressions.

- **Information status.** The possible values for this attribute are `discourse-old`, `discourse-new` and `non-referring`. The value of Kappa is $\kappa = 0.80$.

We have observed in the confusion matrix that the most common disagreement is between the values `new` and `non_referring`.

- **Basic annotation of the markable.** That is the annotation performed before the annotators begin with the annotation of anaphora. The possible values are `discourse-new`, `segment-antecedent` (for discourse deixis), `phrase-antecedent`, and for non referring NPs `expletive`, `quantifier`, `predicate`, `coordination` and `idiom`. The Kappa value is $\kappa = 0.79$.
  The most common disagreement is between the tags `discourse-new` and `predicate`.

- **Semantic type.** The value of Kappa for this feature is κ= 0.85.
  The most common disagreements observed in the confusion matrix are between the categories abstract and concrete and between the categories GPE and organization.

The second set of studies measures the **reliability of the annotated anaphoric links**. First of all we have carried out a study for the selection of antecedent of all anaphoric links annotated in the experiment.

As we have mentioned earlier, empty pronouns in the subject position of the sentence and clitics attached to the verb are two phenomena that appear with a relevant frequency in Italian texts. We have carried out separate studies to test the reliability of the annotation scheme for these phenomena.

- **Link to the antecedent:** The value of Kappa for the annotation of links from markables tagged as `old` to the immediate antecedent is κ = 0.88.

- **Antecedent of clitics:** The value of Kappa for the annotation of links from markables realized as incorporated clitics to the immediate antecedent. $\kappa = 0.84$.

- **Antecedent of empty pronouns.** The value of Kappa for the annotation of links from empty pronouns to the immediate antecedent. κ = 0.93.

4.6 Conclusions

In this chapter I have explained how the annotation instructions have been applied to annotate English and Italian data, how the annotators are trained

and which are the main difficulties in the annotation process.

The annotation of the corpora can be considered as reliable, and the Italian corpus have been used for Semeval.

In the following chapter I present a set of experiments in which I use the annotate data to train and evalute models using features extracted from both corpora.

Chapter 5

Use of the corpora for anaphora resolution

In this chapter I present the results of the use of the produced data for anaphora resolution, and I compare it with the results obtained with other corpora for English and Italian.

In section 5.1 I present BART a modular anaphora resolution toolkit hat have be used for the experiments.

Section 5.2 presents the set of features of the baseline proposed in (Soon, Lim, and Ng, 2001). This baseline is used to test the usability of the annotated corpora for anaphora resolution. The results have been compared with pre-existing corpora, like ICAB for Italian or MUC-7 and ACE-02 for English.

Section 5.3 shows the impact of the manual annotation of MIN_IDs in the resolution of anaphora in the English corpus, and compares the results with the performance of the system in the same corpus without MIN_IDs. For Italian a similar strategy uses the automatically predicted heads.

In Section 5.4 I present some experiments in which we use hand crafted features in the English and Italian corpora.

Section 5.5 enumerates some of the problems of the annotation scheme for anaphora resolution.

Finally section 5.6 summarizes the conclusions of the chapter.

5.1 Tool and classifiers

For the experiments I use the Baltimore Anaphora Resolution Toolkit[1] (BART, (Versley et al., 2008) (Broscheit et al., 2010)). Bart is a highly flexible toolkit based in five main modules as shown in figure 5.1. The modules are:

1. Preprocessing module.
2. Mention factory. extraction of mention pairs (mention factory)
3. Feature extractor
4. Language plugin
5. Learning and classification modules.

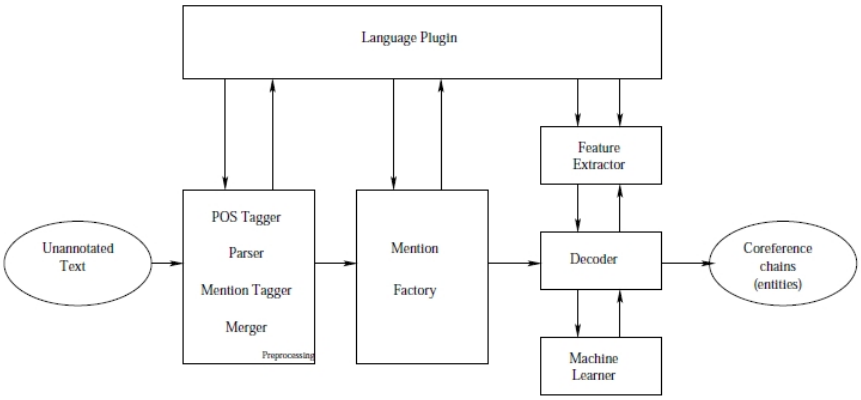

Figure 5.1: Architecture of BART

5.1.1 Preprocessing

Preprocessing consists of marking up noun chunks and named entities, as well as additional information such as part-of-speech tags and merging these information into markables that are the starting point for the mentions used by anaphora resolution.

The toolkit comes with interfaces to several chunkers, parsers or named entity recognizers. For the experiments presented in this section I use a

[1] `http://www.bart-coref.org`

pipeline very close to Soon et al's setup for chunking, named-entity recognition and merging of the information to create markables (Soon, Lim, and Ng, 2001).

In a second step, the mention-building module uses the markables from this layer to create mention objects. Mention objects are grouped into equivalence classes by the resolution process and a coreference layer is written into the document, which can be used for detailed error analysis.

5.1.2 Feature Extraction

The toolkit's default resolver is a reimplementation of the mention pair resolver described in (Soon, Lim, and Ng, 2001).

Each pair of anaphora and candidate is represented as a `PairInstance` object, which is enriched with classification features by feature extractors, and then handed over to a machine learning-based classifier that decides, given the features, whether anaphora and candidate are coreferring or not. I have used this reimplementation as the baseline in the experiments discussed in the following section.

Feature extractors are realized in the toolkit as separate classes, allowing for their independent development.

The set of feature extractors to be used in an experiment is specified in a declarative fashion in an `XML` description file as in figure 5.2, which allows for straightforward prototyping and experimentation with different feature sets.

5.1.3 Language plugin and language specific components

The language plugin module is accessible from each component of BART and handles all the language specific information.

The main language specific components are:

- **English.** The system uses a preprocessing pipeline that integrates the output of the Berkeley Part of Speech Tagger and Parser (Petrov and Klein, 2007) and the Stanford Named Entity Recognizer (Finkel, Grenager, and Manning, 2005).

```
<?xml version="1.0" encoding="UTF-8"?>
<coref-experiment>
<system type="soon">
<classifiers>
<classifier type="maxent" model="idc0"
options="** **"/>
</classifiers>
<extractors>
<!-- general info about antecedent -->
<extractor name="FE_MentionType_Features"/>
<!-- agreement features -->
<extractor name="FE_Gender"/>
<extractor name="FE_Number"/>
<!-- specialized features for aliases / appositive
constructions -->
<extractor name="FE_Alias"/>
<extractor name="FE_Appositive"/>
<!-- string matching features -->
<extractor name="FE_StringMatch"/>
<extractor name="FE_SentenceDistance"/>
</extractors>
</system>
</coref-experiment>
```

Figure 5.2: Experiment description of BART

- **Italian.** The main language specific components described with more detail in (Poesio, Uryupina, and Versley, 2010) are:

  - **Preprocessing:** The text is preprocessed with TextPro (Pianta, Girardi, and Zanoli, 2008), tool that annotates the tokens with sentence boundaries, part of speech tags and named entities. The output of TextPro is merged with syntactic trees produced by the Malt Parser (Nivre et al., 2007). The parse trees are used as well to compute the heads of the markables.
  - **Aliasing:** The plugin implements a feature to cover Italian aliasing patterns, like hand crafted designators for persons and companies, abbreviation, etc.

5.1.4 Learning Modules

The toolkit includes interfaces to Weka and a number of other machine learning packages. For the experiments we use MaxEnt (Phillips, Dudík, and Schapire, 2004).

5.1.5 Evaluation metrics

- MUC (Vilain et al., 1995).
- CEAF (Luo, 2005).
- Link based evaluation: This metric evaluates the links between anaphora and immediately previous antecedent. For instance, if in the gold standard we have a link between two mentions m_a and m_b ($m_b >> m_a$) and the answer of the system is a coreference chain $m_b >> m_x >> m_a$, then precision and recall are both 0. If the answer of the system is $m_b >> m_a$ (correct) and $m_x >> m_a$ (wrong), then precision is 0.5 and recall 1.

5.2 Baseline features

I use a common baseline for English and Italian, the set of features defined by (Soon, Lim, and Ng, 2001) as implemented in BART and summarized in Table 5.1.

Feature	Value	Description
Distance feature		
DIST	integer	the distance in sentences between m_i and m_j
NP type features		
I_PRONOUN	bool	1 if m_i a pronoun
J_PRONOUN	bool	1 if m_j a pronoun
DEF_NP	bool	1 if m_j a definite NP
DEM_NP	bool	1 if m_j a demonstrative NP
Agreement features		
STR_MATCH	bool	1 if m_i and m_j string match
ALIAS	bool	1 if m_j an alias of m_i
GENDER	bool	1 if m_i and m_j gender match
NUMBER	bool	1 if m_i and m_j number match
SEMCLASS	bool	1 if m_i and m_j match semantically
NUMBER	bool	1 if m_i and m_j number match
PROPER_NAME	bool	1 if m_i and m_j both proper names
Syntactic position		
APPOSITION	bool	1 if m_j in appositive position

Table 5.1: Features used by Soon et al (2001)

5.2.1 Baseline results for English

The baseline results for English are summarized in table 5.2. This results are comparable to the results of the resolution in other corpora as one can see in table 5.3.

	Prec.	Recall	F1
MUC	0.609	0.512	0.557
CEAF-AGGR Φ-3	0.669	0.698	0.683
CEAF-AGGR Φ-4	0.677	0.672	0.717
Link-based	0.477	0.623	0.540
Pronouns	0.486	0.655	0.558
Nominals	0.296	0.436	0.352
Names	0.716	0.817	0.763

Table 5.2: Baseline results for ARRAU

Table 5.3 compares the results obtained by BART over ARRAU with the results obtained over the well known corpora MUC-7 and the NPaper dataset of ACE-02.

I present two different results for the ACE-02 corpus. The first one is the result provided by the system using mentions extracted by the CARAFE

mention tagger, a mention extraction procedure based on an ACE-specific mention chunker (Wellner and Vilain, 2006) that gives much better results than using the standard preprocessing pipeline that extracts all mentions regardless of semantic type. A specialized merger discards any base NP that was not detected to be an ACE mention, so that only ACE-compatible semantic types are considered for coreference resolution. Although this resolution model is closer to the state of the art for resolution of anaphoric relations in ACE-like annotation corpora for English, in order to have comparable results I use the results of the BART pipeline to compare the usability of the corpora for anaphora resolution.

	ACE Carafe	MUC-7	ACE02	ARRAU
MUC	0.618	0.585	0.590	0.557
CEAF-AGGR Φ-3	0.537	0.379	0.393	0.683
CEAF-AGGR Φ-4	0.506	0.206	0.309	0.717
Link-based	0.638	0.594	0.532	0.540
Pronouns	0.686	0.492	0.597	0.558
Nominals	0.355	0.455	0.239	0.352
Names	0.638	0.817	0.784	0.763

Table 5.3: Baseline results (F1) for MUC 7, ACE-02 and ARRAU

As we can see in Table 5.3 the results of the system over the ARRAU corpus are comparable with the results obtained over other corpora. That shows that the English corpus annotated following the proposal of chapter 3 is a usable resource for the anaphora resolution task.

Table 5.3 shows too that the results for the CEAF scorer are very low for MUC and ACE data. The reason are that the annotation of these corpora cover less markables than the extracted by the pipeline. The MUC corpus does not contain annotation for the singleton markables, markables that are identified by the system. On the other side the ACE annotation only doesn't cover the annotation of mentions with entity types outside of a list. Markables with other entity type are identified by the system.

5.2.2 Baseline results for Italian

The baseline results for Italian are presented in Table 5.4. I provide two different results, gold mentions (henceforth LMC-Gold) and system extracted markables (henceforth LMC-System). Both results will be used later to an-

alyze the impact of hand annotated features.

	Gold mentions			System mentions		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.429	0.486	0.456
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.638	0.607	0.622
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.711	0.635	0.671
Link-based	0.681	0.505	0.580	0.446	0.498	0.470
Pronouns	0.527	0.516	0.521	0.520	0.520	0.520
Nominals	0.659	0.432	0.522	0.254	0.377	0.303
Names	0.820	0.695	0.752	0.656	0.630	0.642

Table 5.4: Baseline results for LMC

The results are compared in table 5.5 with results obtained with the ICAB corpus, a corpus annotated with an scheme that follows closely the annotation style provided by the ACE annotation scheme. That demonstrates that the usability of the Italian corpus annotated following the annotation scheme presented in this thesis for anaphora resolution.

	ICAB	LMC-Sys	LMC-Gold
MUC	0.494	0.456	0.619
CEAF-AGGR Φ-3	0.557	0.622	0.798
CEAF-AGGR Φ-4	0.560	0.671	0.869
Link-based	0.556	0.470	0.580
Pronouns	0.452	0.520	0.521
Nominals	0.421	0.303	0.522
Names	0.741	0.642	0.752

Table 5.5: Baseline results (F1) for ICAB and LMC

5.3 Annotation of MIN_IDs

In this section I compare the results given by BART for data annotated with MIN_IDs and without them.

The MIN_IDs for the English data have been manually annotated. In the actual phase of the annotation of the LMC corpus the annotation of MIN_IDs

has not been completed. I use automatically produced MIN_IDs produced by the pre-processing pipeline.

	MIN IDs annotated			without MIN IDs		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.539	0.460	0.496
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.480	0.489	0.485
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.449	0.484	0.466
Link-based	0.477	0.623	0.540	0.414	0.623	0.498
Pronouns	0.486	0.655	0.558	0.490	0.691	0.573
Nominals	0.296	0.436	0.352	0.290	0.471	0.359
Names	0.716	0.817	0.763	0.551	0.773	0.644

Table 5.6: ARRAU: Use of MIN IDs

As one can see in the results for the ARRAU corpus shown in table 5.6 and the LMC corpus shown in table 5.7 the use of hand annotated MIN_IDs raise the overall performance of the system for all the metrics. The difference is in particular relevant for the results of the CEAF scorer.

As expected the principal gain of the use of MIN_IDs is in the resolution of names. The reason is that names are markables whose boundaries are frequently difficult to be correctly identified with a parser. In consequence the system has difficulties to align correctly system markables with gold markables, and considers erroneous the markables with non aligned boundaries.

In the resolution of pronouns and nominals the recall decreases for both corpora more than what the increase of precision could compensate.

	MIN_IDs			No MIN_IDs		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.404	0.464	0.432
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.555	0.528	0.541
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.600	0.533	0.565
Link-based	0.446	0.498	0.470	0.413	0.556	0.474
Pronouns	0.520	0.520	0.520	0.491	0.566	0.526
Nominals	0.254	0.377	0.303	0.237	0.431	0.306
Names	0.656	0.630	0.642	0.559	0.647	0.600

Table 5.7: LMC-System: Use of MIN_IDs

5.4 Use of hand annotated features

In this section I present the impact of the use of the following hand annotated features on the results of the evaluation.

1. Gold gender matching.
2. Gold number matching.
3. Gold semantic type matching.
4. Antecedent is subject
5. Grammatic function matching.

5.4.1 Use of hand annotated gender

As one can see in table 5.8 the use of hand annotated gender has a positive effect in the ARRAU corpus for all the evaluation metrics with the exception of CEAF-Φ-4. The improvement is relevant for both, precision and recall.

	Baseline			Gold gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.623	0.543	0.581
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.672	0.702	0.687
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.679	0.753	0.714
Link-based	0.477	0.623	0.540	0.509	0.640	0.567
Pronouns	0.486	0.655	0.558	0.519	0.700	0.596
Nominals	0.296	0.436	0.352	0.326	0.447	0.377
Names	0.716	0.817	0.763	0.738	0.822	0.778

Table 5.8: Results for ARRAU. Use of hand annotated gender

The positive impact of this feature is similar for the LMC corpus in the setting that uses gold markables (table5.9).

In the evaluation using system extracted markables shows a inverse tendency. Table 5.10 shows a better performance only for pronouns. For the other types of mentions the use of the hand annotated information increases the recall of the system, but at the expense of decreasing precision.

	Baseline			Gold gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.541	0.754	0.630
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.806	0.808	0.807
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.923	0.832	0.875
Link-based	0.681	0.505	0.580	0.723	0.504	0.594
Pronouns	0.527	0.516	0.521	0.554	0.522	0.537
Nominals	0.659	0.432	0.522	0.700	0.432	0.534
Names	0.820	0.695	0.752	0.877	0.690	0.772

Table 5.9: Results for LMC-Gold. Use of hand annotated gender

	Baseline			System gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.481	0.441	0.460
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.608	0.578	0.592
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.641	0.619	0.630
Link-based	0.446	0.498	0.470	0.396	0.545	0.459
Pronouns	0.520	0.520	0.520	0.538	0.588	0.562
Nominals	0.254	0.377	0.303	0.213	0.444	0.288
Names	0.656	0.630	0.642	0.609	0.652	0.630

Table 5.10: Results for LMC-System. Use of hand annotated gender

5.4.2 Use of hand annotated number

Table 5.11 shows that the use of hand annotated number for anaphora resolution in the ARRAU corpus has a marginal positive impact for the link based evaluation. The improvement is more relevant in the resolution of names. This result is similar to the result obtained for the LMC corpus with the gold markables setting.

The evaluation results for the LMC-System are summarized in table 5.13. They show a positive impact of the use of hand annotated number for the link based evaluation. The positive impact affects not only the overall results, but the different mention types. The other scorers report contradictory results, with a positive impact for the MUC scorer and negative impact for the CEAF scorer.

	Baseline			Gold number		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.600	0.516	0.555
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.671	0.700	0.685
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.680	0.759	0.717
Link-based	0.477	0.623	0.540	0.487	0.621	0.546
Pronouns	0.486	0.655	0.558	0.496	0.650	0.563
Nominals	0.296	0.436	0.352	0.302	0.438	0.357
Names	0.716	0.817	0.763	0.734	0.815	0.772

Table 5.11: Results for ARRAU. Use of hand annotated number

	Baseline			Gold number		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.549	0.720	0.623
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.797	0.800	0.798
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.909	0.833	0.869
Link-based	0.681	0.505	0.580	0.684	0.507	0.582
Pronouns	0.527	0.516	0.521	0.534	0.522	0.528
Nominals	0.659	0.432	0.522	0.663	0.432	0.523
Names	0.820	0.695	0.752	0.818	0.699	0.754

Table 5.12: Results for LMC-Gold. Use of hand annotated number

5.4.3 Use of hand annotated semantic type

The use of hand annotated semantic type in the ARRAU corpus presented in table 5.14 has a positive impact in the resolution of pronouns, what results in slightly better values for the overall link based evaluation and for the MUC score. The impact on the CEAF scores and on the performance of the system in the resolution of nominals and proper names is just marginal.

The results for the LMC corpus with gold markables are summarized in table 5.15. The only positive effect of the use of the gold semantic types is the reported by the MUC scorer. In the link based evaluation the only improvement is in the resolution of nominals.

5.4.4 Use of the grammatic function

The Italian dataset has not been manually annotated with information about grammatical role of the mentions. To compute the grammatical function of the mentions in the LMC corpus I use two different information sources:

	Baseline			Gold number		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.473	0.534	0.502
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.630	0.599	0.614
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.684	0.611	0.645
Link-based	0.446	0.498	0.470	0.484	0.541	0.511
Pronouns	0.520	0.520	0.520	0.521	0.584	0.551
Nominals	0.254	0.377	0.303	0.301	0.434	0.355
Names	0.656	0.630	0.642	0.716	0.652	0.683

Table 5.13: Results for LMC-System. Use of hand annotated number

	Baseline			Gold semantic type		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.657	0.550	0.599
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.669	0.699	0.684
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.672	0.758	0.713
Link-based	0.477	0.623	0.540	0.506	0.664	0.574
Pronouns	0.486	0.655	0.558	0.541	0.785	0.640
Nominals	0.296	0.436	0.352	0.355	0.472	0.405
Names	0.716	0.817	0.763	0.652	0.800	0.718

Table 5.14: Results for ARRAU. Use of hand annotated semantic type

1. Information about grammatical role of the head of the markable assigned by the MALT parser.

2. Manual annotation of empty subject for verbal markables.

Antecedent is subject

The feature works as follows:

(1) Compute the grammatical role of the antecedent.

(2) If the antecedent is the subject of a sentence, return 1.

(3) Otherwise return 0.

The results for English summarized in table 5.17 show a marginal impact for all the scorers. The main contribution of the implementation of this feature is the impact in the resolution of pronouns.

	Baseline			Gold semantic type		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.570	0.698	0.628
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.793	0.795	0.794
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.892	0.834	0.862
Link-based	0.681	0.505	0.580	0.657	0.522	0.582
Pronouns	0.527	0.516	0.521	0.454	0.686	0.512
Nominals	0.659	0.432	0.522	0.690	0.444	0.541
Names	0.820	0.695	0.752	0.765	0.706	0.734

Table 5.15: Results for LMC-Gold. Use of hand annotated semantic type

	Baseline			Gold semantic type		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.449	0.310	0.366
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.452	0.429	0.440
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.527	0.479	0.451
Link-based	0.446	0.498	0.470	0.282	0.515	0.364
Pronouns	0.520	0.520	0.520	0.714	0.591	0.647
Nominals	0.254	0.377	0.303	0.090	0.355	0.144
Names	0.656	0.630	0.642	0.704	0.645	0.673

Table 5.16: Results for LMC-System. Use of hand annotated semantic type

The results for the LMC corpus show a similar tendency in the experiment performed with gold markables reported in table 5.18. In the experiment realized with system markables and reported in table 5.19 the performance in the resolution of pronouns is lower than the performance for the baseline.

Grammatic function matching

The feature extractor is implemented as follows:

(1) Compute the grammatical function of the anaphora.

(2) Compute the grammatical function of the antecedent.

(3) If if both have the same grammatical function return 1.

(4) Otherwise return 0.

	Baseline			Feature		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.613	0.512	0.558
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.670	0.700	0.684
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.676	0.762	0.716
Link-based	0.477	0.623	0.540	0.479	0.628	0.543
Pronouns	0.486	0.655	0.558	0.484	0.676	0.571
Nominals	0.296	0.436	0.352	0.295	0.436	0.352
Names	0.716	0.817	0.763	0.712	0.815	0.760

Table 5.17: Results for ARRAU. Feature: Antecedent is Subject

	Baseline			Ante subject		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.548	0.725	0.624
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.802	0.804	0.803
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.913	0.835	0.832
Link-based	0.681	0.505	0.580	0.689	0.507	0.584
Pronouns	0.527	0.516	0.521	0.511	0.536	0.523
Nominals	0.659	0.432	0.522	0.672	0.434	0.527
Names	0.820	0.695	0.752	0.848	0.690	0.761

Table 5.18: Results for LMC with gold markables. Feature: Antecedent is Subject

Table 5.20 shows a positive impact of this feature in the resolution of pronouns and increases the precision in the resolution of nominals. But it doesn't have any relevant impact in the overall statistics.

The impact of this feature over the Italian data is not significant as reported in tables 5.21 and 5.22.

5.5 Problematic cases

As one can see in table 5.4 there is a high difference between the results of the setting in which the system uses gold markables and the setting in which the system uses system markables. Although there is a lower recall in the resolution of nominals and proper names, the higher difference is in the precision.

The use of MIN_IDs to facilitate the alignment of the system markables

	Baseline			Gold gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.433	0.498	0.464
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.641	0.610	0.625
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.715	0.636	0.673
Link-based	0.446	0.498	0.470	0.453	0.487	0.474
Pronouns	0.520	0.520	0.520	0.499	0.514	0.506
Nominals	0.254	0.377	0.303	0.271	0.382	0.317
Names	0.656	0.630	0.642	0.657	0.632	0.645

Table 5.19: Results for LMC with system markables. Feature: Antecedent is Subject

	Baseline			Feature		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.609	0.512	0.557	0.614	0.515	0.560
CEAF-AGGR Φ-3	0.669	0.698	0.683	0.670	0.700	0.685
CEAF-AGGR Φ-4	0.677	0.672	0.717	0.676	0.762	0.716
Link-based	0.477	0.623	0.540	0.484	0.631	0.548
Pronouns	0.486	0.655	0.558	0.496	0.669	0.570
Nominals	0.296	0.436	0.352	0.295	0.436	0.352
Names	0.716	0.817	0.763	0.742	0.814	0.776

Table 5.20: Results for ARRAU. Feature: Grammatical function matching

with the gold markables helps to improve resolution of names, but its contribution to the resolution of nominals is not relevant neither for English with hand annotated MIN_IDs (table 5.6) nor for Italian with automatically predicted MIN_IDs (table 5.7).

The task of identifying correctly the markables becomes more difficult in the case of discontinuous markables. Discontinuous markables are markables that doesn't match with syntactic trees or with the output of a name entity recognizer. The identification of this markables require the implementation of additional mechanisms to handle with coordinated noun phrases.

As we can see in table 5.23 if we substitute the discontinuous markables by continuous markables the change in performance of the system is not really relevant. The reason is the low amount of this kind of markables, between a 0.5% and a 7% of the total of markables as reported in section 4.5.3 of chapter 4.

	Baseline			Gold gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.546	0.715	0.619	0.544	0.715	0.618
CEAF-AGGR Φ-3	0.796	0.799	0.798	0.795	0.798	0.796
CEAF-AGGR Φ-4	0.908	0.832	0.869	0.908	0.831	0.868
Link-based	0.681	0.505	0.580	0.680	0.504	0.579
Pronouns	0.527	0.516	0.521	0.526	0.511	0.518
Nominals	0.659	0.432	0.522	0.650	0.432	0.519
Names	0.820	0.695	0.752	0.840	0.691	0.758

Table 5.21: Results for LMC with gold markables. Feature: Grammatical function matching

	Baseline			Gold gender		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.428	0.492	0.458
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.640	0.609	0.624
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.714	0.635	0.672
Link-based	0.446	0.498	0.470	0.452	0.496	0.473
Pronouns	0.520	0.520	0.520	0.518	0.521	0.520
Nominals	0.254	0.377	0.303	0.261	0.375	0.308
Names	0.656	0.630	0.642	0.657	0.630	0.643

Table 5.22: Results for LMC with system markables. Feature: Grammatical function matching

Another difficult task consists of the identification and resolution of empty subjects in the Italian dataset. Nominal markables and verbs with attached clitics might be detected by tools like parser and morphological analyzer. But in the case of empty subject the tools don't give any additional information to the verb that allows to identify it as a markable for the annotation. The most appropriate solution seems to be the implementation of an extra identifier and resolver for empty pronouns.

We have seen that the use of hand annotated information about semantic types can contribute to improve the resolution of pronouns and nominals. But the usability of the annotation of the semantic type of pronouns to train a model is at the moment not an easy task. For instance, the personal pronoun *we* appear often in similar constructions and as subject of the same verbs, but with different semantic types, mostly organization - what often

	Discontinuous mark.			No discontinuous mark.		
	Prec.	Recall	F1	Prec.	Recall	F1
MUC	0.429	0.486	0.456	0.453	0.477	0.465
CEAF-AGGR Φ-3	0.638	0.607	0.622	0.644	0.609	0.626
CEAF-AGGR Φ-4	0.711	0.635	0.671	0.705	0.640	0.671
Link-based	0.446	0.498	0.470	0.429	0.514	0.468
Pronouns	0.520	0.520	0.520	0.492	0.550	0.519
Nominals	0.254	0.377	0.303	0.246	0.404	0.306
Names	0.656	0.630	0.642	0.658	0.631	0.644

Table 5.23: LMC-System: Use of discontinuous markables

happens in the ARRAU WSJ data - and person.

5.6 Conclusions

In this chapter I have presented the use of the annotated dataset for anaphora resolution.

The results obtained with the pipeline based in (Soon, Lim, and Ng, 2001) shows results comparable with the obtained in well known corpora like MUC-7 or ACE-02 for English and ICAB for Italian.

I have discussed the annotation of MIN_IDs, and showed that they can help to improve the resolution of some types of mention, especially the named entities. But the annotation of MIN_IDs is not enough to help in the identification of nominal markables.

I have presented how the use of hand annotated features can affect the resolution of anaphora and different types of mentions. The first observation is that the use of hand annotated morphosyntactic agreement features can be useful to improve the performance of anaphora resolution models for both the English and the Italian data.

Finally I have enumerated cases that are difficult to be resolved, like empty subjects or discontinuous markables.

Chapter 6

Conclusions

I have presented an annotation scheme proposal for the annotation of corpora for anaphora resolution with the aim of defining a set of features applicable for different languages. This scheme has a **higher coverage** than other annotation schemes. It is based in **linguistic criteria** and agreement studies show that the annotation scheme is reliable. The annotated data has been confronted with other datasets in order to demonstrate the **usability** of the produced resources for anaphora resolution.

6.1 Coverage of the annotation

The annotation scheme increases the coverage of other annotation schemes. It gives instructions for the annotation of all kinds of noun phrase, without to constraint the annotation only to several entity types like the ACE like annotation schemes.

It provides instructions to annotate all the noun phrases, including non referring nominal expressions and singleton markables. These markables were not always annotated in the other corpora. For instance the annotation scheme of the MUC corpus provide instructions to annotate only noun phrases that are part of a coreference chain, or the ACE annotation scheme doesn't provide instructions to annotate coordinated noun phrases.

The coverage is also higher than the annotation proposal of MATE and the previous version of ARRAU. The annotation scheme includes instructions to identify as markables nominals and names in premodifier position and non restrictive relative pronouns.

The scheme provides instructions to annotate other kind of anaphoric relations, like the discourse deixis and the bridging descriptions.

Finally the scheme gives instructions to annotate other linguistic features than coreference chains, like morphosyntactic agreement, semantic type or grammatical function, annotation that is not covered in most of the corpora.

6.2 Linguistic criteria

The instructions give a clear account to distinguish different kinds of semantic relations between semantic objects realized as nominal expressions. It provides semantic consistent instructions to annotate identity relations, predication and non-identity anaphoric relations.

A second aspect is the definition of reference, that allows to identify which nominal expressions do not refer to entities in the real world.

The last linguistic relevant issue that I would like to recall here is the treatment of the modification. The annotation gives a criteria to includes all the pre- and post-modifiers in the markable. That motivates the inclusion of discontinuous markables in the annotation.

6.3 Usability of the annotated resources

The annotated English and Italian data has been used for anaphora resolution, and the results have been compared with MUC-7 and ACE02 for English and ICAB for Italian. All these corpora have been used for competitions in coreference resolution.

Finally I have realized experiments with the hand crafted features. The results shows that there are cases in which the impact of the hand annotated data increases the performance of the anaphora resolver. That open the possibility of the use of the data not only to train models of anaphora resolution, but to use the hand annotated data for the training of models to introduce automatically the relevant linguistic information in the decoding.

The Italian data has been selected as dataset for the coreference resolution task at SemEval 2010.

6.4 Further work

The annotation of the datasets is a work in progress. We are extending the manual annotation of MIN_IDs to the full Italian dataset.

The LMC-Wikipedia dataset will increase the volume of annotated data in a 100%, and new datasets will be incorporated to the corpus, like data from the newspaper L'Adige and data from blogs.

The English dataset will be expanded too. The new features of the annotation scheme will be included in other datasets, like Gnome, the Vieira-Poesio Corpus or the Pear Histories. Finally a new batch of trees from the Penn Treebank will be converted in the MMAX format and annotated using the instructions presented in this thesis.

References

Aone, Ch. and D. McKee. 1993. A language-independent anaphora resolution system for understanding multilingual texts. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 156–163, Morristown, NJ, USA. Association for Computational Linguistics.

Aone, Chinatsu and Scott Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *ACL*, pages 122–129.

Bosco, C., V. Lombardo, D. Vassallo, and L. Lesmo. 2000. Building a treebank for italian: a data-driven annotation scheme. In *Proceedings of the LREC-00*, Athens. Greece.

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162, Morristown, NJ, USA. Association for Computational Linguistics.

Broscheit, Samuel, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 104–107, Uppsala, Sweden, July. Association for Computational Linguistics.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Chafe, Wallace L., editor. 1980. *The Pear stories : cognitive, cultural, and linguistic aspects of narrative production.* Ablex Pub. Corp., Norwood, N.J. :.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chinchor, Nancy. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference (MUC-7).*

Clark, Herbert H. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, TINLAP '75, pages 169–174, Morristown, NJ, USA. Association for Computational Linguistics.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Gross, Derek, James F. Allen, and David R. Traum. 1993. The trains 91 dialogues. Technical report, University of Rochester. Computer Science Department., Rochester, NY, USA, July.

Harabagiu, Sanda M. and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the sixth conference on Applied natural language processing*, pages 142–149, Morristown, NJ, USA. Association for Computational Linguistics.

Heeman, Peter A. and James F. Allen. 1995. The trains 93 dialogues. Technical report, University of Rochester. Computer Science Department., Rochester, NY, USA, March.

Hendrickx, Iris, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for dutch. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2008`.

Hinrichs, Erhard W., Sandra Kübler, and Karin Naumann. 2005. A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 13–20, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Hirschman, Lynette and Nancy Chinchor. 1997. Muc-7 coreference task definition. In *Proceedings of the 7th Message Understanding Conference*.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, NL.

Krippendorff, Klaus, 80. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage Publications, Inc, December.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535–561, December.

LDC, Linguistic Data Consortium, 2004. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1.*

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Magnini, B., E. Pianta, Ch. Girardi, M. Negri, L. Romano, M. Speranza, and R. Sprugnoli. 2006. I-cab: the italian content annotation bank. In *Proceedings of the LREC-06*, Genova, Italia.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.

McCarthy, Joseph F. and Wendy G. Lehnert. 1995. Using decision trees for conference resolution. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1050–1055, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mengel, A., L. Dybkjaer, J.M. Garrido, U. Heid, M. Klein, V. Pirrelli, M. Poesio, S. Quazza, A. Schiffrin, and C. Soria. 2000. MATE Deliverable D2.1. MATE Dialogue Annotation Guidelines.

Mitkov, R. and C Barbu. 2000. Improving pronoun resolution in two languages by means of bilingual corpora. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, Lancaster, UK.

Mitkov, R. and M. Stys. 1997. Robust reference resolution with limited knowledge: high precision genre-specific approach for english and polish. In *In Recent Advances in Natural Language Processing (RANLP-97*, pages 74–81.

Müller, Ch. and M. Strube. 2006. Multi-level annotation of linguistic data with mmax2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt a. M., Germany.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kbler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 2(13).

Passonneau, Rebeca. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA).

Petrov, S. and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proc. HLT-NAACL*.

Phillips, Steven J., Miroslav Dudík, and Robert E. Schapire. 2004. A maximum entropy approach to species distribution modeling. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 83, New York, NY, USA. ACM.

Pianta, E., Ch. Girardi, and R. Zanoli. 2008. The textpro tool suite. In *Proceedings of LREC-08*.

Poesio, M. and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of LREC-08*, Marrakech, Marocco.

Poesio, M., R. Delmonte, A. Bristot, L. Chiran, and S. Tonelli. 2004. The venex corpus of anaphora and deixis in spoken and written italian. Manuscript.

Poesio, Massimo. 2004a. Discourse annotation and semantic annotation in the gnome corpus. In *DiscAnnotation '04: Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Morristown, NJ, USA. Association for Computational Linguistics.

Poesio, Massimo. 2004b. The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

Poesio, Massimo and Ron Artstein. 2005. Annotating (Anaphoric) Ambiguity. In *Proceedings of the Corpus Linguistics Conference*.

Poesio, Massimo, Olga Uryupina, and Yannick Versley. 2010. Creating a coreference resolution system for italian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources*

and Evaluation (LREC'10), Valletta, Malta, may. European Language Resources Association (ELRA).

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Comput. Linguist.*, 24:183–216, June.

Pradhan, Sameer S., Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 446–453, Washington, DC, USA. IEEE Computer Society.

Recasens, M., Ll. Márquez, E. Sapena, A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (Semeval 2010)*.

Recasens, Marta and M. Mart. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, pages 1–31.

Rodriguez, K.J., F. Delogu, Y. Versley, E.W. Stemle, and M. Poesio. 2010. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta.

Soon, W. M., D. C. Y. Lim, and H. T. Ng. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4), December.

Steger, J. M. and E. W. Stemle. 2009. The architecture for unified processing of web content. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, Donostia-San Sebastian. Basque Country.

Strube, Michael, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 312–319, Morristown, NJ, USA. Association for Computational Linguistics.

van Deemter, Kees and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Versley, Yannick, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12, Columbus, Ohio, June. Association for Computational Linguistics.

Vilain, M., J. Burger, J. Aberdeen, C. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52, Los Altos, CA, EUA, 6-8 de Novembro. Morgan Kaufmann.

Webber, Bonnie Lynn. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.

Wellner, Ben and Marc Vilain. 2006. Leveraging machine readable dictionaries in discriminative sequence models. In *In Language Resources and Evaluation Conference, LREC 2006*.