Annotating Bridging Anaphors in Italian: in Search of Reliability.

Tommaso Caselli* and **Irina Prodanof†**

*†Dipartimento di Linguistica, Università degli Studi di Pavia,
C.so Strada Nuova, 65 27100 Pavia
*Dipartimento di Linguistica “T. Bolelli”, Università degli Studi di Pisa,
Via Santa Maria, 36 56100 Pisa
†Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche Pisa
Via Moruzzi, 1 56124 Pisa
{cobweb80@yahoo.it } {irina.prodanof@ilc.cnr.it}

Abstract

The aim of this work is the presentation and preliminary evaluation of an XML annotation scheme for marking bridging anaphors of the form “definite article + N” in Italian. The scheme is based on a corpus-study. The data we collected from the evaluation experiment seem to support the reliability of the scheme, although some problems still remain open.

1. Introduction

Bridging relations are of major importance in establishing and maintaining textual coherence and, at the same time, the most challenging problem in anaphora resolution. The correct resolution of bridging structures is important for various applications like Open-Domain Q.A. and for Information Extraction and Retrieval systems.

Kleiber (1999) defines bridging anaphora as a “type of indirect textual reference whereby a new referent is introduced as an anaphoric not of but via the referent of an antecedent expression” [Kleiber 1999: 339]. This means that they cannot be resolved on the basis of string matching and thus require the reader to built up a “bridge” by using common-sense inference mechanism.

A trend in linguistic theories, which had counterparts in computational frameworks, tends to emphasise the idea that full definite noun phrases (FDNPs) are a matter of the global discourse focus i.e. they are used to retrieve a referent which is no more accessible or to construct a conceptual representation which uniquely identifies a referent. On the contrary, empirical studies provided evidence to Sidner’s (1979) hypothesis that bridging FDNPs are different from other occurrences of anaphoric FDNPs, since, in the process of identification of their anchors (or antecedents), they are more sensitive to the local focus. The claim we make is that the agreement among annotators for bridging FDNPs can be improved by applying a strict set of linguistic tests and a preference ranking. In addition to this, a centering-based analysis of each discourse segment should reduce problems of multiple anchoring.

2. The Corpus Study

The scheme has been developed on a corpus-study on 17 randomly chosen articles for a total of 10,000 words, from the Italian newspaper “*il Sole-24 Ore*”, a corpus used in the SI-TAL Project, the syntactic-semantic Treebank of Italian.

The texts considered contain a total number of 1412 full definite noun phrases (FDNPs) of the form “definite article + (possessive) + N”, which represent 31.54% of all the occurrences of FDNPs in the corpus. Each newspaper article was first read entirely, and only after it was divided into segments of five sentence windows which is an arbitrary strategy to give an account of the local focus of the text i.e. the most probable place to look for anchors for bridging FDNPs.

In the classification exercise we have used an operational device such as processing requirements since when a FDNP is encountered in a text/discourse can be reduced to one of these four cases:

- it is used to pick up an entity mentioned before in the text, which, in our experiment, could be either directly or indirectly realized;
- it is not mentioned before, but its interpretation depends on , is based on, or is related in some way to an entity already present in the text/discourse (directly or indirectly realized);
- it is not mentioned before and is not related to any previous mentioned entity, but it refers to something which is part of the common shared knowledge of the writer and reader;
- it is self-explanatory or it is given together with its own identification.

These four types of FDNPs use reflect the classes of Direct Anaphora, Bridging and First Mention, respectively. The same operational device i.e. processing requirements, was used for the analysis and classification of bridging anaphors.

The classification task has led to the identification of 6 main classes of FDNPs (Table 1 below) and 5 subclasses of bridging anaphors (Table 2 below).

To maintain the quality of annotation a special set of heuristics has been created, in particular for identifying First Mention FDNPs and to disambiguate the role of modification. One of the most interesting results from the data in Table 1 is the high number of First Mention definites (58.61%). The percentage rises to 61.15% if we include the class of Possessives. These results of the class First Mention represent further empirical support for Löbner’s (1985) theory of definiteness.

CLASS	NUMBER OF ITEMS	PERCENTAGE
First Mention	833	58.61%
Possessives	36	2.54%
Direct Anaphora	170	12.03%
Bridging	299	21.17%
Idiom	25	1.62%
Doubt	49	3.47%
Total	1412	100%

Table 1- Classes of FDNPs.

To maintain the quality of annotation a special set of heuristics has been created, in particular for identifying First Mention FDNPs and to disambiguate the role of modification. One of the most interesting results from the data in Table 1 is the high number of First Mention definites (58.61%). The percentage rises to 61.15% if we include the class of Possessives. These results of the class First Mention represent further empirical support for Löbner's (1985) theory of definiteness. The relative few instances of Direct Anaphora FDNPs (only 12.03%) is due both to language specific reasons, since Italian allows zero anaphora on verb subject, and in part to stylistic reasons linked to the linguistic subdomain of newspaper articles. Support to this hypothesis is provided by comparison with other empirical studies conducted in English. The class of Bridging represents the 63.88% (299/469) of all anaphoric FDNPs, suggesting that bridging is a more productive cohesive strategy in Italian with respect to other languages, i.e. English.

The data in Table 2 provide empirical evidence to the theoretical debate by supporting the claim that lexico-encyclopaedic knowledge and discourse structure play a primary role in the process of resolution of bridging anaphora. The results also suggest a preference order for the different sources of bridging anaphora: lexical semantic relations are preferred over the use of common sense inferencing and background knowledge i.e. pragmatics, which is preferred over discourse structure. In addition to this, we have found out that 119/221 (53.48%) of the anchors identified have been either backward-looking centres (Cbs) or preferred centres (Cps) of previous sentences and that at least 70% of the anchors can be found either in the current or in the immediate previous sentence, providing thus further empirical evidence to Sidner's (1979) claim. Particular attention should be placed on the Event class whose anchor is not an NP but a VP. The remaining anchors are realized by other elements in the forward-looking centre (Cf) list. The ranking of the elements in the Cf list according to the thematic role suggests a preference of anchors in oblique position over indirect object.

The data in Table 2 are only in part comparable to those found by Poesio et alii (2004b) and Poesio (2003), since in these works the search was restricted to cases of merological bridging, and the corpus used was very different. In addition, no support to the claim that the first mention entity of a previous sentence is likely to be the anchor was found. This means that:

CLASS	NUMBER OF ITEMS	PERCENTAGE
Lexical	119	39.79%
Event	18	6.02%
Rhetorical Relation	27	9.03%
Discourse Topic	26	8.69%
Inferential	109	36.45%
Total	299	100%

Table 2- Subclasses of bridging anaphors.

- knowledge of the local focus is necessary but not sufficient to determine the anchor of a bridging description;
- a preference ranking in the search of probable anchors is very useful, since it increases the precision of the process of resolution of the bridging descriptions, provided that there are ways to check the plausibility of the proposed solution.

3. Architecture of the Scheme

In the design of this XML annotation scheme for bridging FDNPs the principle of standoff annotation and the data presented in Section 2 are the starting points for its development.

The first step is represented by the definition of the markables i.e. the class of entities between which the relations to be annotated can occur. The textual elements in which we are mostly interested are FDNPs of the form "definite article + N" in anaphoric relation and those parts of a text which provide antecedents. In Table 3 (below) the tags used in the scheme and their attributes are illustrated:

TAGS	**ATTRIBUTES**
<ne>	*CAT, GEN, NUM, PER, GF, LF_TYPE, ANAPHORIC, BRIDGING, TITLE*
<ve>	*SEMTYPE, ARGSTR*
<seg>	*TYPE, ANAPHORIC*
<link>	*REL, LOOKBACK*
<ante>	*CENTERING*

Table 3- Tags and their attributes.

We introduce three tags for the markables:

- <ne> : for Nominal Expressions;
- <ve> for Verbal Expressions; and
- <seg> used for the indirect realization of pronouns and for clitics.

One of the most innovative aspect concerning the markables is represented by the <ve> tag which is assigned to the verb, including the auxiliaries, since as we found in the corpus-study some bridging FDNPs have a Verb Head as anchor and not a nominal expression (the class of Event in Table 2). The attribute

ARGSTR (argument structure) is very important since it helps in identifying a part of Event bridging.

To distinguish the markable <ne>, we introduce the attribute *CAT*, which allows us to differentiate between different subtypes of nominal expressions such as proper name, definite NP and so on. In addition, attributes like *PER*, *GEN*, *NUM* and *GF* capture other grammatical features of interest like person, gender, number and grammatical function.

Interesting attributes for the tags <ne> and <seg> are *ANAPHORIC*, *BRIDGING* and *TITLE*. They all have Boolean values (yes/no). In particular, the first reflects both Fraurud's (1990) claim and the corpus results that the only distinction which can be marked reliably is that between first mentions and subsequent mentions. The attribute BRIDGING represents an attempt to mark instances of bridging anaphora without trying to identify the type of relations, which is mainly responsible for disagreement between annotators. The attribute TITLE is strictly related to the corpus. We claim that the entities in the titles of newspapers represent meta-textual objects, thus we consider titles as a meta-level of the discourse/text. This attribute is responsible for the identification of bridging referring to Discourse Topic.

The anaphoric expression and its relation with the anchor are coded by the element <link> which is a structured element, i.e. it has an embedded element, <ante> which identify the anchor of the anaphoric expression.

The attribute *REL* in the <link> tag is used to annotate the anaphoric relation. Its values are extended with respect to GNOME and VENEX, and includes *ident*, *subset*, *poss*, *elem*, *frame*, *event*, *title* and *underspec*. It should be noted that the assignments of values to this attribute is strictly linked to the value of *ANAPHORIC* and *BRIDGING*. The measure of salience is provided by the attributes *LOOKBACK* and *CENTERING*.

The presence of the attributes *ANAPHORIC* and *BRIDGING* in the <ne> tag and the value *underspec* in *REL*, offer a solution both to the problem of ambiguity in the identification of bridging FDNPs and to the consequent creation of multiple paths. This can be done by restricting the number of previous sentences where the probable anchor should be searched (as done in the corpus-study, Section 2) and providing a preference order which states that whenever there is an identity relation this should be preferred, even if the anaphoric expression can enter a bridging relation with another available anchor.

We expect an improvement of the K value for bridging anaphors recognition between $0.70 \leq K \leq 0.80$ and a higher value in the identification of the anchors with respect to previous studies (Vieira 1998,Vieira-Poesio 2000).

4. Overview of the Evaluation.

The evaluation has been conducted by the authors and two university students, one of them graduated in Linguistics, henceforth annotators A1, A2 and A3. A collection of three articles from the corpus containing 267 FDNPs was first classified by the authors and next the two subjects were asked to perform the same task.

The classes are modified with respect to those presented in Section 2 Table 1, since the class of Idiom has been considered as part of the class First Mention. The two subjects were provided with a manual containing a strict set of rules on how to conduct the annotation. The rules instructed the two subject to resolve conflicts, and thus reducing ambiguity and disagreement, according to a decision tree based on a series of linguistic tests (i.e. syntactic structure, role of modification, presence of special predicates and so on and so forth) and a preference ranking i.e. to choose a class with higher preference. The ranking was First Mention > Direct Anaphora > Bridging. In addition to this, for every FDNP marked as Bridging a specific preference ranking based on Centering Theory for the identification of the correct anchor had to be used. The ranking was Cbs > Cps > object(s) of the Cf in Oblique position > object(s) of the Cf in Indirect Object position. Annotator A2 was previously given a brief training to familiarise with the task.

4.1 General Evaluation

The results collected are shown in Table 4 below:

CLASS	A1	A2	A3
First Mention	140 (52.43%)	138 (51.68%)	132 (49.44%)
Possessives	15 (5.62%)	15 (5.62%)	15 (5.62%)
Direct Anaphora	41 (15.36%)	47 (17.60%)	46 (17.23%)
Bridging	69 (25.84%)	67 (25.09%	65 (22.48%)
Doubt	2 (0.75%)	0	9 (3.37%)
Total	267 (100%)	267 (100%)	267 (100%)

Table 4- Annotators' classification of FDNPs.

As the table indicates, all annotators assign approximately the same percentage of FDNPs to each class. The per-class agreement measure was computed. The rates of agreement per each class thus obtained are: First Mention 93.4%, Direct Anaphora 93.2%, Bridging 82.08% and Possessives 100%.

There were 188 cases of complete agreement among annotators on the classification (70.41%): 103 on First Mention, 37 on Direct Anaphora, 33 on Bridging and 15 on Possessives.

To measure the agreement in a more precise way the Kappa statistic coefficient was computed. The overall coefficient of agreement among the three annotators (A1, A2 and A3), excluding the class of Doubt, is $K = 0.73$ (for 257 FDNPs).

The first important result is the relative high agreement among annotators. The value of the K for all FDNPs allows us to claim that the scheme thus developed and its rules are reliable. In addition, the rates of the per-class agreement for the class of Bridging are very good and higher with respect to those obtained by Vieira (1998) in both her experiments (59% in Experiment 1 31% in Experiment 2).

The cases of disagreement are due, mainly, to the fact that the classes are not mutually exclusive and the wrong application of a test may lead to an incorrect classification (example 1); the FDNP *il World Trade Center* has been classified as Bridging by A3 because of the presence of the word *torri* (towers) in the previous sentence but according to the annotation rules this is not allowed since all first occurrences of proper names must be marked as First Mention:

1) *L' attentato fece 6 morti e mille feriti nelle* ***torri*** *di New York .// Condannati all' ergastolo negli Usa i terroristi del* World Trade Center .

The other reason for disagreement is due to a wrong segmentation of the text into the five sentence window, so that it can happen that cases of Direct Anaphora are marked as First Mention or Bridging.

4.2 Evaluating Bridging

The results we obtained for the class of Bridging are quite different. We have seen that the per-class agreement is 82.08%. On the other hand, the value of the *K* coefficient, is as low as *K*= 0.58. Although this value is not satisfying, and very far from the one expected, it is much much higher than the one found by Vieira–Poesio (2000) where *K*= 0.24.

As already stated in Section 4.1, the reason for the disagreement among the annotators is mainly due to the wrong application of the linguistic tests. To confirm our intuition, we have computed the *K* only between annotators A1 and A2, who was given little training before completing the task. As expected the value of the *K* coefficient improves and it is as high as *K*= 0.71, which is in line with our expectation. However, a new evaluation with different annotators previously trained must be conducted.

The per-class agreement and the *K* coefficient evaluate the agreement on classification of the uses of FDNPs. When evaluating the class of Bridging, and in particular our claim of a positive correlation between centering-based analysis of the sentences which form a discourse segment and agreement on the anchors, we need a way of assessing agreement on the identification of a discourse antecedent. To do this we considered the rate of agreement on the anchors over agreement on the proper classes. In our study 33 cases of FDNPs were classified by all three coders as Bridging. For these 33 cases they have identified the same anchor for 26 of them, with 78% agreement on the antecedent.

5. Conclusion and Future Work

The results presented in the previous Section show a good reliability of the annotation scheme for classifying the uses of the FDNPs.

The results obtained for the class of Bridging are not as good as expected and a second evaluation experiment will be necessary, in order to confirm the intuition that to improve agreement between annotators a training is required. The second notable result is very positive, in that it confirms that to reduce multiple paths in the identification of anchors salience and local focus play a major role.

Finally, since the data from this preliminary investigation are quite positive and reliable, the next step will be the implementation of a real annotation tool and the development of methods for automatically recognize bridging FDNPs.

References

Asher, N and A. Lascarides.1998, Bridging, in *Journal of Semantics* vol. 15 (1): 83-113.

Caselli T. 2005 *Lexical anaphora: a corpus-based model and an XML annotation scheme for bridging anaphora in Italian*, M.A. Thesis, University of Pavia

Chierchia, G.1995, *Dynamics of Meaning: anaphora, presuppositions and the Theory of Grammar*, University of Chicago Press, Chicago.

Fraurud, K.1990, Definiteness and the Processing of Noun Phrase in Natural Discourse in *Journal of Semantics*, vol. 7 (4): 395-433.

Gardent, C. et alii.2003, Which bridges for bridging FDNPs?, in *Proceedings of the Fourth International Workshop on Linguistically Interpreted Corpora*, Budapest, Hungary.

Grosz, B. et alii.1995, Centering: A framework for modelling the local coherence of discourse, in *Computational Linguistics*, vol. 21 (2): 202-225.

Kleiber, G.1999, Associative anaphora and part-whole relationship: the condition of alienation and the principle of ontological congruence, in *Journal of Pragmatics*, vol. 31: 339-362.

Korzen, I.2003, Anafora associativa: aspetti lessicali, testuali e contestuali, in *Atti del XXXIV Congresso Internazionale di studi della Società di Linguistica Italiana (SLI). Firenze 19-21 ottobre 2000*, Maraschino, N. and T. Poggi Salani (eds)Bulzoni, Roma.

Poesio, M.2003, Associative FDNPs and salience, in *Proceedings of the EACL Workshop on Computational Treatments of Anaphora*, Budapest.

Poesio, M., R. Stevenson, B Di Eugenio and J. Hitzeman.2004a, Centering: A Parametric theory and its instantiations in *Computational Linguistics*, vol. 30 (3).

Poesio M., R Metha, A. Maroudas and J. Hitzeman. 2004b, Learning to Resolve Bridging References, in *Proceeding of ACL-04.*

Prince, E.1981, Toward a Taxonomy of given-new information, in *Radical Pragmatics*, P. Cole (ed), Academic Press, New York.

Vieira R.1998 *Definite Description Processing in Unrestricted Text*, Ph.D. Thesis, University of Edinburgh.

Vieira, R. and M. Poesio.2000, An Empirically-Based System for Processing FDNPs, in *Computational Linguistics*, vol. 26 (4): 539-593.