Mental Models Theory and Anaphora

Guido Boella and **Leonardo Lesmo**
Dipartimento di Informatica and Centro di Scienza Cognitiva
Università di Torino
email: {guido, lesmo}@di.unito.it

Abstract

We argue that anaphora cannot be resolved at the level of the formal language representing meaning, but, rather, by making direct reference to the *extension* of the sentences. Johnson-Laird's mental models theory provide the tool for coping with extensional representations in a cognitively plausible way.

Introduction

Anaphoric expressions are traditionally viewed as substitutes for more complex linguistic expressions which have already occurred earlier in the text. Anaphora has proven difficult to analyze at a purely syntactic level, so that structural approaches like DRT [10] or semantic ones like Dynamic Semantics [4] cope with this problem by enriching the formal language used to build or to represent the meaning of sentences.

We believe that the limit of these approaches is that they have chosen the wrong level of representation for dealing with anaphora: we will show that it is necessary to make direct reference to *extensional* representations of meaning. In particular, the representation of the context should put at disposal the elements of the situation, which anaphors can refer to, instead of hiding them behind quantified expressions.

However, extensions can possibly be infinite or too large to be dealt with directly. But there is a proposal which uses extensional representations of finite and limited size, and which has been shown to be cognitively plausible, i.e., the *mental models theory* of [9]. Johnson-Laird has used mental models in order to explain how people reason without having to resort to formal logic. Inferences are performed by manipulating extensional representations of sentences which are composed of a finite number of elements and relations: *"a mental model represents the* extension *of an assertion, i.e., the situation it describes, and the recursive machinery for revising the model represents the* intension *of the assertion, i.e., the set of all possible situations it describes."* (p.100)

In [8]'s words: *"mental models theory is a psychological theory of language processing and reasoning. The theory provides a framework within which more detailed accounts of the component processes of comprehension [...] such as anaphora interpretation [...] and reasoning can be developed, [...] Mental models theory assumes that comprehension results in the construction of representations of situations in the real world [...] These models are finite and computable, and they are constructed incrementally, with the model so far acting as part of the context for interpreting the current text.* (p.20)
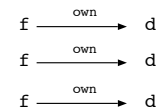
A simple preliminary example illustrates our solution. In the following sentences, the acceptability is guaranteed just for the pair where the (intended) antecedent (*a donkey*) and the pronoun (*they*) do not agree in their syntactic number:

(1) *Every farmer owns a donkey. *It is pink.*
(2) *Every farmer owns a donkey. They are pink.*

When the second sentence in each discourse is interpreted, it produces a mental model which must be *integrated* with the preceding one: a referent must be found for the anaphoric expressions. If we examine in the Figure below how the first sentences of the two pairs are represented in mental models theory, we see that the problem is easily solved. The mental model contains a finite number of tokens (placeholders for individuals,

here farmers *f* and donkeys *d*) and relations among tokens (the arrows labeled with own). Given the model above, which donkey, out of the represented ones, can we relate to the singular *it*, appearing in the second sentence of (1)? The problem of identifying the referent appears to be the same as in: (3) *I have three sisters. *She is blonde* where we have to choose one referent out of three candidates. One is given no (or not enough) information to identify the antecedent (among the three sisters) denoted by *she*. On the contrary, the *they* pronoun in (2) can be interpreted as referring to the set of donkeys appearing in the model, due to its plural syntactic number.

The mental model building algorithm

First of all, the sentence undergoes a syntactic and semantic interpretation process that produces a semantic network (see [6], [11] and [2] for details on the network representation). Then, following the proposal by [9], that *"a propositional representation can be used as the input to a procedural semantics that constructs mental mod-*

els", a mental model representing the meaning of the sentence is built.

The network representation

For the present purposes, we will describe briefly only the mechanism of Distributivity Ambiguity Spaces (DAS) which deals with the possible distributive readings of an NP (see [11] for details).

The nodes of the network can be simple or DASs. The latter correspond to plural NPs, and they were introduced to deal with the distinction between *collective* and *distributive* readings of predicates: each DAS includes two subnodes *Set* and *Indiv*.

In case of (4) *Three men lifted three tables*, if the subject NP is given a reading as a set, the men are seen as being jointly involved in the act of lifting tables. Viceversa in the individual reading of the subject, each man executed a separate lifting act. If the tables are interpreted as a set too, they were lifted all together (perhaps they were stacked). On the contrary, if they are interpreted as individuals, the men lifted them one at a time. The four combinations of *Set, Indiv* readings for the subject and the object do not cover all possibilities. In fact, it may happen that, for the *Indiv* reading of the subject, there exist just three tables, and each man lifted one of them (three individual lifting acts); or that each man lifted three tables (possibly, but not necessarily, the same three tables; 9 different tables could be involved), so that nine individual lifting acts have been executed. Or, in the *Set* reading of the object, the three men lifted three different stacks of tables (so, we have two more readings, for a total of 6) The extra readings (see Figure 1) are accounted for by means of a mechanism other than the DAS described above (but independently motivated, see [11]), i.e. by the presence of DEP-ON (dependent on) arcs. They are similar to Skolem functions in first order logic, and were introduced for representing quantifier scoping. Each node which is not universally quantified can be specified to be dependent on another 'plural' node. For instance, in *Every farmer owns a donkey*, the most natural reading is where each farmer owns a different donkey, so that the particular donkey 'depends on' the particular farmer.

Mental models

In order to use a more unambiguous version of the framework with respect to the 'diagrammatic' original version of [9], we refer to the formalization of mental models provided by [1].

According to [1], a model is triple $< T, R, A >$, where T is a (non-empty) bi-dimensional matrix of tokens, R is a set of relations on T, and A is a set of annotations. For dealing with some interpretations, more than one model can be required.

A token is either a model or an element. An element is a pair $< S, A >$ where S is a symbol from a given vocabulary and A is a set of annotations; the vocabulary consists of named individual entities (john for the proper name *John*) and generic entities belonging to some category (c_i for cars, f_i for farmers, etc.).
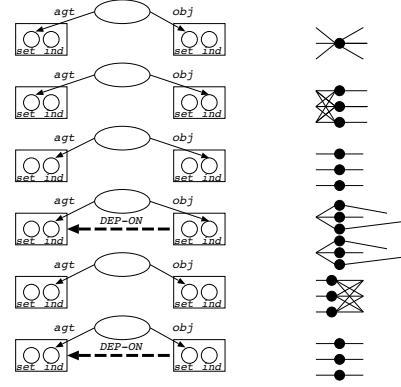

Figure 1: Six readings of (4).

A relation is an ordered sequence $< r, x_1, \ldots, x_n, A >$ where r is a relation symbol, $x_1, \ldots, x_n$ are tokens in T and A is a set of annotations. Annotations are "*the propositional enrichment of the analogical structure of the model*" [1]. In particular, the "not" annotation applies to any feature of the models. For models and relations, a negation means that they are not the case; for entities, that they are absent in a model. The "..." annotation means that the model can be further extended.

[1] consider relations such as 'above', 'faster' and two special relations "connected with" (CW) and "never connected with" (NCW). The CW relation forms an individual by connecting two of its properties. The NCW one states that two properties cannot hold for the same individual. Usually the two relations are used to represent the meaning of, respectively, *all humans are mortals* and *most lawyers are not poor*. With respect to [1]'s framework, we introduce an extension for what concerns the NCW relation. In fact, NCW is originally meant to apply only to unary predicates such as being humans or mortal. We introduce a version of the NCW relation relativized to a predicate rel, NCW(rel). In fact, the "not" annotation of a relation means that the relation is not true of the given entities involved in the relation. In $<$ faster, john, bill, $\{not\} >$, the negation does not concern the existence or not of the two individuals John and Bill, which are introduced as existing entities. But this is not sufficient to represent the meaning of a sentence like *John does not have a car*: since the phrase *a car* inside a negation does not introduce or refer to an entity in the model, the meaning of the sentence cannot be represented by the negation of the 'have' relation: in fact, a relation as $<$ have, john, $c_1, \{not\} >$ does not express the fact that from the model it is not possible to infer that there is a car. Rather, this annotated relation expresses the fact that there is a car in the model and John is not its owner.

What we need is something similar to the interpretation of the sentence *no lawyer is a crook* from which is not possible to infer that there is some crook in the model. The model of this sentence in [1] is not represented by a negation of some predicate 'is' but with the NCW rela-

tion discussed above: $< \text{NCW}, \text{lawyer}_i, \text{crook}_i >$. Analogously, for interpreting *John does not have a car* we introduce a NCW(have) predicate, which not only expresses the negation of the "have" predicate, but which also does not assert the existence of any car (see the Figure below).

In $<$NCW(have), john, c_1, $\emptyset >$, cars, as in $<$ NCW, bicycle_i, c_i, $\emptyset >$ (*no bicycle is a car*), are kept separate from the other entities in the model: they cannot play the role of antecedents of pronouns.

For what concerns the treatment of logical connectives, we stick to the proposal of [1].

From the network to the mental model

The model constructing procedure takes as input an existing mental model (representing the context) and the network representation of the new sentence (still associated with the syntactic tree): the newly constructed model is *integrated* with the existing ones by overlapping identical tokens and finding referent tokens for anaphoric expressions.

The process starts from the non-dependent entity nodes of the network which derive from the interpretation of NPs (i.e. NPs without exiting DEP-ON arcs), and proceeds with the other NPs, according to the (partial) order imposed by (reversed) DEP-ON arcs. After that, all co-references are solved. For instance, in (5) *Every farmer who owns a donkey beats it*, *every farmer* is processed first, then *a donkey* and, the pronoun *it* which depend on the subject NP.

More precisely, given a context M composed by a model $< \text{T}, \text{R}, \text{A} >$, we have that a network W is interpreted as a new model $< \text{T}', \text{R}', \text{A}' >$, in the following way:

1. Each non-dependent entity node in the network W deriving from the interpretation of an NP is treated separately:

   (a) If the entity node is represents an NP which is a proper noun (e.g., *John*), an individual token (e.g., john) is introduced in the matrix T of the model M; if that token is already present in the model, the two tokens are identified.

   (b) If the NP is a quantified Noun (e.g., *every farmer*), a set of distinct tokens $\text{F} = \{x_1, \ldots, x_n\}$ representing the denotation of the noun is added to the context matrix T; depending on the quantifier Q, a subset of them, Q(F), will be selected for linking to other tokens by the relation where the NP occurs as an argument (selecting the whole set in case of *every* and *all*, a proportioned subset of it in case of *most*, etc). The annotation A of the model can be augmented with a "**...**", since, depending on the quantifier, more tokens could be added to the matrix T or the set $Q(F)$ could be revised (e.g., if Q="some", $\mid Q(F) \mid$ could be initially 2 or 3, but it can be increased in case of necessity, as in the standard

treatment of syllogism in [9]).

A special case, as in the mental models theory of [9], is represented by the quantifier *no*: its meaning is represented by selecting all the tokens F representing the denotation of the noun it quantifies (Q(F)=F); but when the relation rel involving the NP is introduced, it is interpreted as negated either in the sense of a NCW(rel) relation or in the sense of being annotated as negated. As an example, in *no farmer owns a donkey* the owning relation, is transformed in a NCW(own) relation which keeps apart all the farmers from the set of donkeys.

(c) If the NP is an indefinite such as *a car*, two cases are possible according to the presence of a negation and the role played by the NP in the main predicate:[1]

- If the NP is the subject of the verb or it appears in a non-negated relation, a single new token representing a car is added to the matrix T of the model and annotated as "...", since it does not convey any uniqueness presupposition.
- If the NP appears in a negated predicate and it is not the subject of the predicate rel, some tokens representing the denotation of the noun $\text{F} = \{x_1, \ldots, x_n\}$ are introduced in T and appear in a NCW(rel) relation to keep them separate from the other tokens of the model.[2]

(d) If the entity in the network W is the interpretation of a definite NP or a definite pronoun, then an antecedent must be searched for in the mental model constructed so far; according to the number, one or more tokens existing in the model are sought in T to act as the potential referents: further, the set of relations R must satisfy the description provided by the NP. This kind of unification, however, cannot be accomplished with items which are linked to other ones only by a NCW(rel) relation in which they appear in a non-subject role $\{t_i \mid \exists \text{ rel}, x_1, \ldots, x_n \ (< \text{NCW(rel)}, x_1, \ldots, t_i, \ldots, x_n, \emptyset > \in \text{R} \land i \neq 1)\}$, i.e., these items are implicitly assumed as 'non existing' in the model. Moreover, if the set of possible referents $\text{X} = \{t_1, \ldots, t_n\}$ is composed of a subset of tokens which occur in relations with other tokens and a subset of tokens which are unrelated:
$\{t_i \mid \exists \text{ rel}, \text{A}, x_1, \ldots, x_n \ (< \text{rel}, x_1, \ldots, t_i, \ldots, x_n, \text{A} > \in \text{R}) \} \cup \{t_j \mid \neg \exists \text{ rel}, \text{A}, x_1, \ldots, x_n \ (< \text{rel}, x_1, \ldots, t_j, \ldots, x_n, \text{A} > \in \text{R}) \}$
then only the former set can be considered by the uni-

---

[1]Note that *John does not love a girl in his office* where the indefinite is a *specific* one (see [10]) and the speaker could identify a unique referent for it, is not covered by this rule.

[2]This treatment of indefinites is justified also from a linguistic point of view. As [10] notice, the negation of a verb must be interpreted as having an inner scope which does not include the subject of the verb, otherwise sentences as *someone does not like a Porsche* would be true in case there is no people at all. And it finds a similarity in DRT where indefinites inside the scope of a negation are interpreted in a subordinate DRT structure which will not be accessible for the resolution of anaphoric expressions.

fication process (e.g., in the interpretation of *John has many donkeys. They are pink* where the model includes a number of donkeys but only a subset of them is related with John: the pronoun *they* refers only to this subset).

Note that the set of annotations is not constrained to be empty: in fact, it is possible to make reference to a set of entities which is involved in a negated relation as in: (6) *the soldier didn't see some of the enemies. They were hiding in the trees.*

Finally, since a definite pronoun is a *definite* reference, the found referent must be non-ambiguous: if different possibilities exist, then, for pragmatics reasons, the reference fails (see example (3)).

2. If the entity node of the NP np_1 is "dependent on" another node which is built from the NP np_2, its interpretation depends on the one of np_2: this means that, for each token built in correspondence with np_2 the interpretation of np_1 must be repeated according to the rules in 1 described above for non-dependent NPs. In particular, if np_1 is a singular *indefinite* and the corresponding relation is not negated, a new token is introduced for each token associated with np_2; if np_1 is plural, a different set of example tokens is added to the model for each token associated with np_2.

   For example in the distributive interpretation of *Every farmer has a donkey. They beat it*, *they* is unified with the tokens $f_1, \ldots, f_n$ representing farmers, but the interpretation of *it* (which in this reading cannot but be dependent on *them*) is performed for each f_i ($1 \le i \le n$) relatively to the set of tokens $\{ t_j \mid \exists \text{ rel}, x_1, \ldots, x_n$ $(< \text{rel}, x_1, \ldots, f_i, \ldots, t_j, \ldots, x_n, \emptyset > \in R) \}$. In the example, for each i, *it* is unified with the d_i such that $<\text{beat}, f_i, d_i, \emptyset >$.

3. Finally, the tokens are linked by the relations described by the predicates. The number of relations which are introduced depends on the *set* or *individual* interpretation of the DAS of the NPs involved: if an NP is considered as a set, the tokens resulting from its interpretation are included as a whole in the role they play in the relation. Otherwise, each element of the set is introduced in different instance of the relation.

4. As we discuss in the following Section, the interpretation of a sentence which includes logical connectives can result in more than one model. The rule 1 is iterated for each of the clauses in the complex sentence.

   During the interpretation process some of the possible models must be discharged as inconsistent. This is a correct move but it can lead to the refection of the sentence for pragmatic reasons (as in example (11) below). In fact, if the interpretation of a sentence results in a reduced set of models which can be better described by another sentence (that is, its interpretation does not discard any model), then by the Gricean principle of cooperation, the speaker should have used it instead of the one he chose.

5. On the other hand, if the interpretation of the sentence leads felicitously to a set of models, these models be-

come part of the context. When a subsequent sentence is interpreted, its interpretation must be compatible with *all* the models in the context. In particular, if the interpretation of the subsequent sentence produces more than one model, for each model in the context, at least one of the newly constructed models must be compatible (even if not the same one for all the model in the context). Otherwise, the sentence will be rejected (as in example (14) below).

Logical connectives

According to [10] the interplay of anaphora and logical connectives is a fundamental testbed for any theory of language interpretation. Here, the meaning of connectives is expressed by their possible models in [9]'s style. First the implicit models are constructed and if necessary the explicit ones are fleshed out.

Let's start with a simple example involving negation: (7) *John does not own a car. He washes it.*

Since, according to the representation outlined in the previous section, cars are included in NCW(own) relations, no referent can be found in the model for the pronoun *it*: $< \text{T}=\{\{\emptyset, c_1\}, \{\text{john}, \emptyset\},\}\}, \text{R}= \{<\text{NCW(own)}, \text{john}, c_1, \emptyset >, \text{A}= \emptyset \} >$

So, the sentence is not interpretable according to that reading.

An example a bit more complex is: (8) *No farmer has a car. *It is red.*

A sentence like *no farmer is rich* is represented by a NCW relation between farmers and rich people see rule 1.b. In our model, this relation is extended to arbitrary predicates. Hence, the first sentence produces a model where cars appear in the set of *never connected with* entities, so that the interpretation (and failure in integration) is exactly the same as in the previous example: $< \text{T}=\{ \{\{\emptyset, c_1\}\{\emptyset, c_2\}\{\emptyset, c_3\}\}, \{\{f_1, \emptyset\}, \{f_2, \emptyset\}, \{f_3, \emptyset\}\} \}, \text{R}= \{<\text{NCW(own)}, f_1, c_1>, \emptyset >, < \text{NCW(own)}, f_2, c_2>, \emptyset >, <\text{NCW(own)}, f_3, c_3>, \emptyset >\}, \text{A}= \{...\} \}$

On the contrary: (9) *No farmer has a car. They prefer donkeys.* is acceptable, in spite of the negation appearing in the subject NP and of its singular number. In fact, the farmers (appearing as 'existing' entities) are available for integration.

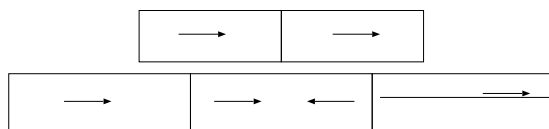If we now consider conjunctions and disjunctions, another interesting anomaly arises:

(10) *John owns a car$_i$ and Fred washes it$_i$*

(11) **John owns a car$_i$ or Fred washes it$_i$*

The syntactic structures are identical but the acceptability is not. In order to explain this fact, [10] introduced an accessibility constraint at the structural level: "*no disjunct of a disjunctive condition is accessible from any other*".

The mental model representation of a conjunction involves the inclusion in the same model of the conjoined sentences. So, no problem arises with (10), since the referent for *it* can be found in the same model where the second conjunct must be integrated. Compare the unacceptability of **John does not have a Porsche and Fred washes it.*

On the contrary, a disjunction A ∨ B requires the construction of two *separate* models (one with A and one with B). So, in the second model of (11) there is no available referent for the pronoun (B, i.e., *Fred washes it*). But when a difficulty (such as the impossibility of understanding a sentence) occurs in a mental model, the model can be manipulated and fleshed out; in principle, when applied to the second model (B), this process could produce two alternatives. In the first one John owns a car (A and B), while in the second one he does not (not A and B, i.e., *John does not have a car and Fred washes it*). So, it seems that the first extension could solve the problem: John, in fact, owns a car and Fred washes it:

The lower part of the figure shows the three resulting models: the second one (A and B) includes the first (A) and the third (not A and B) is discharged since Fred cannot wash a car which does not exist (see the * in the third box).

However, it seems that from any disjunction at least two distinct models must be constructed, and that none of them must be included in the other: otherwise, the common part of the two models would be necessarily true and according to Grice, the speaker should not have used a disjunction to express such a meaning (see rule 4 of the interpretation algorithm).

An example that supports the previous analysis is the acceptability of the following sentence, as the reader can easily test: (12) *John does not own a car or he washes it*. The second model (the one of *he washes it*) can be fleshed out with the negation of the first disjunct (not A and B, i.e., John does own a car and Fred washes it): after this extension, the resulting model puts at disposal the required referent for the pronoun. We are left with two different models, equivalent to the [10]'s interpretation of the example, i.e *John does not own a car* or *John owns a car and he washes it*. In [10] this result is obtained by copying the negation of the first conjunct in the second DRS: such a rule, however, presupposes that disjunctions in natural language are always interpreted a exclusive disjunctions.

It is interesting to note that Dynamic Semantics [4], in order to explain this kind of examples, has to introduce a new class of anaphora, *E-type*.

The last connective to be considered is implication: (13) *If John owns a car, he washes it*.

This sentence involves two models (A and B, and not A to be further extended): in the first one, John owns a car and washes it, while in the second one he does not own any car. But, if the sentence is followed by (14) **It is a Porsche*, the pronoun of the second sentence can find a referent only in the first model in the context, while no antecedent is accessible in the second one (where the

car, in fact, is connected with a NCW relation). As prescribed by rule 4 of the algorithm, a new sentence must be integrated with all the pre-existing models in the context, otherwise it is unacceptable.

An analogous reasoning explains the oddity of **If John does not own a car, Fred washes it*. Moreover, so called *examination sentences* as (15) *no students will be admitted to the exam unless they have registered four weeks in advance* can be dealt with by interpreting the conjunction *unless* as an *if not* or a stronger *exclusive or*.

Plurals

A number of interesting anaphoric phenomena are related to *plurals*. The first situation concerns a plural pronoun referring to a set of singular antecedents which occur in the model: (16) *Mary met John. They talked*.

Even if in mental models the first sentence introduces separately two tokens, the rule 1.d can cope with these case as if they were introduced simultaneously as by the NP *two people*. So it can resolve the anaphora without explicitly *summing* the antecedents to form a plural discourse referent, as DRT does.

Quantifier phrases such as *many of the farmers* do not always introduce the referents with which subsequent pronouns will be co-referential. For example, those pronouns refer to sets that have to be constructed from explicit information in the text. Here, quantifiers introduce in the model a set of tokens which pronouns can refer to in the subsequent discourse: (17) *Susan has found every book which Bill needs. They are on his desk*.

To resolve *they* we need only the set of tokens introduced in the model by the analysis of the first sentence. The right subset of books is identified in the model thanks to rule 1.d (see discussion below). In DRT, in contrast, a new discourse referent is constructed via an *abstraction* rule which copies the content of the DRSs introduced by the previous sentence.

In some approaches, definite NPs and pronouns inside the scope of a quantifier are considered like bound variables in a logical system. In (18) *every waiter wants customers to give him large tips* the pronoun does not seem to refer to any particular entity, while it does in (19) *John wants customers to give him large tips*. DRT in order to deal with both cases in a uniform way introduces the notion of discourse referent which does not correspond directly to any individual in the world, while providing antecedents for the pronouns.

In contrast, mental models theory allows unifying both cases, since quantifier phrases, in an extensional representation introduce sets of entities in the model.

Not all definite pronouns following quantifiers behave like bound variables, in particular, if they appear in a following clause, i.e., outside the quantifier scope: (20) *few congressmen admire Kennedy and they are very junior*. *They* refers to those (few) congressmen who admire Kennedy, even if there is no such an expression referring to them. If the pronoun were interpreted as a variable, the sentence would be equivalent to (21) *Few congressmen admire Kennedy and are very junior*. In [5]'s terms, there

is an *antecedent trigger*, a linguistic expression which introduces the antecedent of the pronoun but it does not have the same referent of the pronoun.

In our model, after the interpretation of the first clause the mental model contains the set of congressmen and a (small) subset of them which are in an "admire" relation with Kennedy. For rule 1.d above, the definite pronoun *they* can be resolved with this subset.

But as it might be expected, quantifiers focus on the subset of the set specified by the head noun. Hence, the unification process must be suitably constrained. In: (22) *Some farmers of this valley own a donkey. They don't like cars*, the pronoun *they* can in principle refer either to the *farmers of this valley* who own a donkey or to the complement set; according to rule 1.d in the interpretation algorithm, if the set of candidate referents can be partitioned in different sets, the pronoun is unified only with the entities which are involved in a relation (of owning a donkey).

The possibility of a plural anaphor resolved against referents described by a singular indefinite is explained by rule 1.d which deals with the interpretation of dependent NPs in distributive readings (see the Figure on first page). In (23) *Every farmer owns a donkey. They are pink* the distributive reading expresses explicitly the plurality of donkeys so that the correct referents are available for the plural pronoun. In contrast, in (24) *Every farmer owns a donkey. *He is a wise man*, the singular definite pronoun *he* cannot be resolved, since we do not have any information to choose one of the farmers (see rule 1.d).

An example slightly more complex is: (25) *Three farmers own a donkey. They beat them*. The latter sentence can be interpreted only as far the second clause is interpreted with two individual readings of the NP without DEP-ON arcs between them (case 3 of Figure 1). In this case the donkeys who form the antecedent of *them* are related each by a different relation with the farmers. Which farmer is selected for relating with the beating relation to a given donkey? as in case of rule 2 the interpretation of an anaphor is performed exactly with respect to the other tokens which are linked to it by some relation. Indeed, in the context is maintained the relation between each farmer and the donkey he owns: hence, the interpretation of the sentence leads to a situation where each farmer beats the donkey he owns, and not a different one (as it happens in some formal models of anaphora).

Finally, plural and singular pronouns can be mixed: (26) *Every farmer owns a donkey. They beat it*. Since *it* in the second sentence is dependent on the subject *they*, the interpretation of the second sentence is parallel to the interpretation of the first: the object (*it*) can be resolved against an antecedent only if it is interpreted as dependent on the subject; according to rule 1.d of the interpretation algorithm, *they* is unified with the set of farmers appearing in the model and, again, since there is an explicit relation (*own*) linking each of them to a donkey, this link is followed to determine the (singular) referent of *it*.

Similarly, the so called *donkey sentence*, (27) *Every farmer who owns a donkey beats it*, is acceptable: the procedure first interprets the subject phrase, thus obtaining, in the wide-scope reading of the universal, a representation where each farmer has at least a donkey; then it extends the representation by searching for a referent through each relation $<$own, f_i, d_i, $\emptyset >$; so, in the distributive reading the sentence, as in the example above, for each farmer, a different referent for *it* is found, i.e. the donkey owned by him. The possible antecedent must satisfy the restriction carried by the number of the singular pronoun (compare **Every farmer who has two donkeys beats it*).

Conclusion

Since mental models are a cognitively plausible theory of human reasoning, they can be also useful in finding an explanation of linguistic phenomena. In [3], we exploited mental models to provide an explanation of lexically triggered presuppositions. In [2] more complex anaphorical phenomena related to the different readings of *donkey sentences* have been coped with in the same framework.

The limit of logical approaches in explaining anaphora is that they exploit representations that are not isomorphic to our conception of the described situation. The necessity of resorting to a referential level in explaining anaphora has been highlighted also by [7]. We followed his suggestion, but going in a different direction, where mental models replace the classical model-theoretic framework to provide a cognitively plausible approach to language interpretation.

References

[1] B. Bara, M. Bucciarelli, and V. Lombardo. Model theory of deduction: A unified computational approach. *Cognitive Science*, 25(6), 2001.

[2] G. Boella, R. Damiano, and L. Lesmo. Beating a donkey: a mental model approach to complex anaphorical phenomena. In *Proc. of European Congress of Cognitive Science of ECCS'99*, Pontignano, 1999.

[3] G. Boella, R. Damiano, and L. Lesmo. Mental models and pragmatics: the case of presuppositions. CogSci99 Conference, 1999.

[4] G. Chierchia. Anaphora and dynamic binding. *Linguistics and Philosophy*, 15:111–183, 1992.

[5] F. Cornish. Antecedentless anaphors: Deixis, anaphora or what? *Journal of Linguistics*, (32):19–41, 1996.

[6] B. DiEugenio and L. Lesmo. Representation and interpretation of determiners in natural language. In *Proc. 10th IJCAI*, pages 648–653, Milano, 1987.

[7] D. A. H. Elworthy. A theory of anaphoric information. *Linguistics and Philosophy*, 18:207–332, 1995.

[8] A. Garnham. *Mental models and the interpretation of anaphora*. Psychology Press, Hove, 2001.

[9] P.N. Johnson-Laird. *Mental Models*. Cambridge University Press, Cambridge, 1983.

[10] Hans Kamp and Uwe Reyle, editors. *From Discourse to Logic*. Kluwer, Dordrecht, 1993.

[11] L. Lesmo, M. Berti, and P. Terenziani. A network formalism for representing natural language quantifiers. In *Proceedings of ECAI-88*, pages 473–478, 1988.