# Models of Anaphora Processing and the Binding Constraints

## 1. Introduction

In cognition-driven models, anaphora resolution tends to be viewed as a surrogate process: a certain task, more resource demanding, is reformatted in terms of an equally effective but simpler task. When some entity happens to be recurrently referred to, the task of first mention interpretation is taken over by a task of anaphora resolution: The search in a larger cognitive space — long-term or semantic memory — is avoided by means of a search in a more restricted cognitive space — short-term or working memory. Anaphora processing is thus usually seen, more or less explicitly, as a case of search optimization motivated by the reduction of the cognitive search space in the semantic processing of referring expressions.

### 1.1. The search optimization rationale

Building further on this search optimization rationale, different types of anaphors — e.g. pronouns vs. definite descriptions, personal pronouns vs. demonstratives, etc. — have been assumed to be resolved against different admissible antecedents from different "sections" of the relevant cognitive search space. While interpretive overhead is reduced by shifting to anaphoric reference, anaphoric interpretation in turn is facilitated by a "divide to conquer" strategy: The search space for finding antecedents for anaphors is "sectioned", each section being searched for in the resolution of anaphors of a given type or class.

As different items in working memory are distinguished from one another given their relative attentional prominence, anaphors of a given type can thus pick up items with a certain degree of, or in a certain range within the hierarchy of, attentional prominence, while anaphors of another type pick items with some other degree of attentional prominence.

Skimming through the literature, one finds different proposals concerning the number of "sections" into which the search space for anaphora resolution is expected to divide. Just a few examples: Authors like Guindon (1985) or Givón (1992) discuss a division, respectively, into two and three "sections", Gundel *et al.* (1993), in turn, proposes a schema that may extend the division up to six "sections", depending on the specific language at stake.

### 1.2. Natural classes of anaphors from natural classes of antecedents

In line with the tenets of the rationale above, different sorts of anaphors — whose antecedent entities are to be found in different "sections" of the search space — are expected to have different sets of admissible antecedent entities.

The strong prediction is thus that anaphors of different types have different, *disjoint* sets of antecedents, a position that can be found e.g. in (Garrod and Sanford, 1982).

Another, weaker but still plausible prediction in this connection is that, if the different sets of antecedents turn out not to be disjoint, they are at least expected to be *successively included* within each other. If we admit that an anaphor is of a given type such that it searches or is sensitive to items with a certain degree of attentional prominence, it is not a contradiction to accept that this anaphor may also be sensitive to items with a higher degree of prominence. This is the intuition behind the approach of Gundel *et al.* (1993, 1998).

The search optimization rationale for anaphora — with the assumed correlation between anaphoric types and attentional prominence of corresponding admissible antecedents — can thus be seen as inducing a delimitation of anaphors into different natural classes. These classes are naturally circumscribed because every anaphor in that class can be resolved against the same class of admissible antecedents.

The point worth stressing then is that this sets up a very interesting line of inquiry: If one succeeds in isolating different classes of admissible antecedents, then we will succeed in isolating natural, cognitively motivated classes of anaphors. This line of inquiry is one of major relevance also because, if one finds such natural classes of anaphors, this represents a piece of empirical support of paramount importance for cognition-driven models of anaphora processing sharing the search optimization rationale mentioned above.

## 2. Antecedent accessibility

A major goal in this line of research involves assuming a suitable scale for the attentional prominence of admissible antecedent entities and finding objective criteria to decide with which part of the scale an anaphor should be put in correspondence with. The pursuing of this goal has been reported at various places in the literature, cf. among others, (Prince, 1981) and (Gundel *et al.* 1993).

### 2.1. Fuzzy delimiters

The scales used to evaluate the attentional status of the cognitive item corresponding to a given anaphor are typically defined by means of a set of keywords, like "familiar", "activated", "evoked", "uniquely identifiable", "brand new", etc. These keywords appear with definitions under the form of examples and a discussion of some cases to which they may apply. The keywords are also associated with a hierarchy, where the relative positioning of each keyword in the scale is defined *vis a vis* the other keywords.

This approach seems to be problematic, in our view, in some crucial aspects. There is not an empirical justification for the number of required keywords, that is of distinct degrees of relevant attentional prominence. Keywords are defined in such a way that the boundaries between the degrees of prominence they are supposed to delimit are not clear. Crucially, and above all, there is no objective criteria to unequivocally decide which point of the scale is a given anaphor in correspondence with.

These shortcomings represent a drawback for the goal of finding empirical support for the search optimization rationale of anaphora resolution. This does not mean, however, that they might be seen as empirical justification to reject the conjectures embodied in such rationale.

Our point in bringing to light the above deadlock is not the dismissing of cognition-driven models of anaphora processing as such. Our line of argument is rather that overcoming this deadlock may involve changing the angle from which the correlation between natural classes of anaphors and search optimization is addressed.

Instead of in first place looking at objective criteria to identify attentional status and then trying to use them to possibly delimit classes of anaphors, we should take into account actual natural classes of anaphors — empirically motivated on the basis of differences in classes of admissible antecedents — and try to clarify their eventual cognitive underpinnings. In particular, we should discuss whether and how such classes may fit into a search optimization rationale for anaphora resolution.

### 2.2. Binding classes

One of the most notorious group of classes of anaphors obtained via grouping of corresponding sets of antecedents are the so called binding classes. Each of these classes contains anaphors that may pick an antecedent from the same set of admissible antecedents. The members of a given class are intensionally characterized as those anaphors that obey a specific binding constraint, with this constraint expressing an objective criterion to categorize anaphors.

Binding constraints delimit the relative positioning of anaphors and their admissible antecedents in grammatical and discourse geometry.[1] From an empirical perspective, these constraints stem from quite robust generalizations and exhibit a universal character, given their parameterized validity across natural languages. From a conceptual point of view, in turn, the relations among binding constraints involve non-trivial symmetry, which lends them a modular nature. Accordingly, they have been considered one of the most robust aspects of linguistic knowledge, usually known as binding theory.

Since their first formulation in (Chomsky, 1980, 1981), the specification of binding constraints has been the focus of intense research in last decades, from which a binding theory of increased empirical adequacy has emerged. Recent developments of (Pollard and Sag, 1994) indicate that there are four of such constraints (vd. Xue *et al.*, 1994, Branco and Marrafa, 1999):

> Principle A
>
> A locally o-commanded short-distance reflexive must be locally o-bound.
>
> *Lee$_i$ thinks* [*Max$_j$ saw himself$_{*i/j}$*].

[1] In the context of our discussion, a noteworthy point is that, in first place, binding constraints correlate the interpretation of anaphors with linguistic structure, not with the attentional status of the cognitive representation of the admissible antecedents. In the framework of centering theory, besides being correlated with the attentional status of the cognitive representation of their admissible antecedents, anaphors are correlated also with discourse and grammatical structure (vd. Walker *et al.*, 1998b). As this theory unfolds its predictions basically for pronouns in anaphoric links across adjacent sentences, no grouping of natural classes of anaphors of the kind we are concerned here is implied by those predictions.

Principle Z
An o-commanded long-distance reflexive must be o-bound.
*[O amigo do Rui$_i$]$_j$ acha que o Pedro$_k$ gosta dele próprio$_{*i/j/k}$.* (Portuguese, Branco and Marrafa, 1999)
[the friend of_the Rui] thinks that the Pedro likes of_he self
'[Rui$_i$'s friend]$_j$ thinks that Pedro$_k$ likes him$_{*i/j}$/himself$_k$.'

Principle B
A pronoun must be locally o-free.
*Lee$_i$ thinks [Max$_j$ saw him$_{i/*j}$].*

Principle C
A non-pronoun must be o-free.
*[Kim$_i$'s friend]$_j$ thinks [Lee saw Kim$_{i/*j}$].*

These constraints are defined on the basis of some auxiliary notions. The notion of *local domain* involves the partition of sentences and associated grammatical geometry into two zones of greater or less proximity with respect to the anaphor. Typically, the local domain coincides with the predication domain of the predicator subcategorizing the anaphor. In some cases, there may be additional requirements that the local domain is circumscribed by the first upward predicator that happens to be finite, bears tense or indicative features, etc.

*O-command* is a partial order under which, in a clause, the Subject o-commands the Direct Object, the Direct Object o-commands the Indirect Object, and so on, following the usual obliqueness hierarchy of grammatical functions, being that in a multiclausal sentence, the upward arguments o-command the successively embedded arguments.

The notion of *o-binding* is such that *x* o-binds *y* iff *x* o-commands *y* and *x* and *y* are coindexed, where coindexation is meant to represent anaphoric links.[2]

### 2.3. Sets of admissible antecedents

As discussed above, the search optimization rationale for anaphora resolution implies some predictions concerning the relations between the different classes of admissible antecedents. These are expected to be either disjoint — strong prediction —, or successively included within each other — weak prediction.

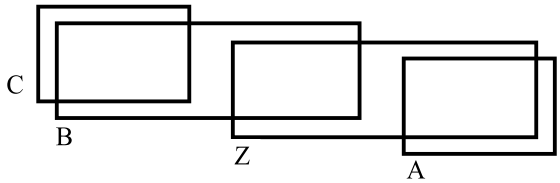
Given the binding classes just presented, we can now check if they support these predictions. For each different binding class we delimit the corresponding set of admissible antecedents and then check out how these sets of admissible antecedents relate to each other. In order to proceed with this test we just have to fix a (non-exempt) position in a generic multiclausal grammatical structure and successively instantiate that position with an anaphor **x** from each of the four different binding classes. This way we will be able to collect the four sets of admissible antecedents and observe what are the relations among them.

Accordingly, if we assume that **x** is an A-anaphor complying with principle A, we see that its admissible antecedents form the set of the local o-commanders of **x**, which we can call the set A. In case **x** is a Z-anaphor, the set Z of its admissible antecedents is made of its o-commanders. When **x** is a B-anaphor, the set B of admissible antecedents contains all the antecedents that are not local o-commanders of **x**. Finally, the set C of the admissible antecedents of **x** when this is a C-anaphor has all the antecedents that are not o-commanders of **x**.

The relations between these four sets are as displayed below. Note that, the exact details of binding constraints may vary from language to language given the different, language specific definitions for the auxiliary notions of *local domain*, etc. Also, some languages may not have instances of all of the four types of anaphors. In this connection, the important point to stress is that, what seems to remain invariant cross-linguistically is the fact that, however the details of binding constraints turn out to be set up in each language, the relations between the corresponding sets of admissible antecedents hold as follows:

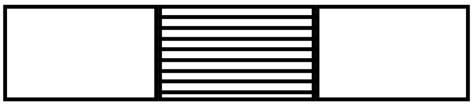
$$A \subset Z \qquad Z \cap B \neq \varnothing \qquad B \subset C$$
$$A \cap B = \varnothing \qquad Z \cap C = \varnothing$$
$$A \cap C = \varnothing$$

Representing these sets of admissible antecedents in graphical terms, we obtain:



[2] There are anaphors that are subject-oriented, in the sense that they only take antecedents that have the grammatical function Subject. Some authors (e.g. Dalrymple, 1993) assume that this should be seen as an intrinsic parameter of binding constraints and aim at integrating it in their definition. In this point we follow (Branco and Marrafa, 1999), where the subject-orientedness of anaphors is argued to be, not an intrinsic feature of binding constraints, but one of the surfacing effects that result from the non linear obliqueness hierarchy associated with some predicators (or to all of them in some languages).

The relations among them can be rendered in a more perspicuous way in 2-D by means of the following Venn diagram:



It is straightforward to check that the possible antecedents of short-distance reflexives are possible antecedents for long-distance reflexives; some possible antecedents of long-distance reflexives are possible antecedents of pronouns; and the possible antecedents of non-pronouns are possible antecedents of pronouns. From another, more general perspective, for a given possible antecedent, it is the case that there are always at least two different types of anaphors that can take it.[3] In any case, what is crucial to note is that the sets of admissible antecedents per anaphor type are not mutually disjoint. They are neither successively included within each other. This does not match either the strong or the weak prediction implied by the search optimization rationale.

An important conclusion thus resulting from this discussion is that the search optimization rationale does not seem to offer a justification of binding classes.

## 3. Looking for other sort of cognitive underpinnings

It is a fact that not all binding constraints, v.z. the ones concerning reflexives, have an impact in terms of constraining the anaphoric capacity of anaphors whose interpretive anchoring may hold across dialogue segments. Be that as it may, if we intend to find some cognitively plausible model for anaphora resolution, such model should support all kinds of anaphors, with all different kinds of anaphoric capacity, including those restricted only to the sentential domain. Accordingly, given the considerably well established empirical grounding of binding constraints, the problems now uncovered cast non negligible doubts on the explicative value of models for anaphora processing of the kind mentioned above.

These considerations should not be seen, however, as implying that cognitively rooted factors (as for instance, among other, attentional prominence associated with recency of mention) do not play an important role in anaphora resolution, at least as preference mechanisms.[4] Nor should it be seen as implying that binding constraints have been proved not to have any cognitive justification.

Instead, these negative results should be seen, in our view, as showing that cognitive underpinnings of binding classes are perhaps entangled in other aspects of cognition. This suggestion is supported by the results from previous work of ours, which may provide interesting hints as to the possible cognitive grounding of binding constraints.

Building on the existence of a fourth binding principle, for long-distance reflexives, Branco and Marrafa, 1999 showed that the binding principles can curiously be arranged into a square of logical oppositions. This stimulated our research on the eventual quantificational character of binding constraints. Adopting Löbner's, 1987 duality criterion for quantification in natural language, and the formal tools he developed for the analysis of phase quantification, we showed in (*self-reference*) that the four binding constraints can actually be seen as the effect of four binding quantifiers. These phase quantifiers are expressed by the nominals of the four binding classes, and quantify over the reference markers organized in the grammatical obliqueness hierarchy.

This proposal lends support to a new understanding of the formal nature of binding constraints, and to interesting explanations of some related issues such as the exemption occurrences and logophoric behavior of reflexives.

Given the space constraints inherent to this extended abstract, presenting a full-fledged account of the empirical support and justification as well of the implications of these results has to be postponed to the presentations of the full paper. This does not hinder, nevertheless, a brief statement of the reasons behind our suggestion

[3] If one considers instead an exempt syntactic position (cf. Pollard and Sag, 1994), a position where the anaphor has no o-commander, then even reflexives have possible antecedents that may also be antecedents of pronouns and non-pronouns.

[4] Since the so called integrative approach to anaphora resolution was set up in late eighties ((Carbonell and Brown, 1988), (Rich and LuperFoy, 1988), (Asher and Wada, 1988)), and its practical applicability extensively checked up, (cf. (Lappin and Leass, 1994), (Mitkov, 1997), (Mitkov, 1998) among others) it became common wisdom in the literature on computational models that factors for determining the antecedents of anaphors divide into filters or constraints, and preferences or heuristics. The first exclude impossible antecedents and help to determine the set of antecedent candidates; the latter help to pick the most likely candidate, that will be proposed as the antecedent. Binding constraints belong to the set of filters.

that these results might hint at a renewed rationale for the cognitive grounding of binding constraints: Given the quantificational structure underlying them, binding constraints, and in particular their apparent universal, cross-linguistic validity, may be seen as a manifestation of quantification, a universal semantic module of natural language, which is arguably expected to be rooted in some cognitive invariant.

## 4. Summing up

In this paper we first pointed out a common assumption underlying cognition-driven models of anaphora resolution that there is a search optimization rationale behind eventual constraints on the antecedents against which anaphors may be resolved This rationale implies some predictions about the existence of natural classes of anaphors. In particular, it implies that the sets of admissible antecedents for each such natural class bear certain relations among them. These sets are predicted either to be disjoint, or at least to be successively included within each other.

Given such relations are not observed for the sets of admissible antecedents corresponding to binding classes, these natural classes of anaphors are not offered any principled explanation by that rationale. Moreover, given the fact that they are some of the most notorious natural types of anaphors defined in terms of classes of their admissible antecedents, this result casts doubts that the search optimization rationale may provide an encompassing cognitive justification for anaphora processing.

When looking at alternatives for some sort of cognitive explanation of binding constraints, recent results on the quantificational nature of these constraints may be seen as interesting hints on a yet to explore line of research: Natural language quantification rests on a duality-based logical structure, which, being a universal semantic module of natural language, is arguably expected to be rooted in some cognitive invariant.

## 5. References

Asher, N. and H. Wada, 1988. A Computational Account of Syntactic, Semantic and Discourse Principles for Anaphora Resolution. *Journal of Semantics*, 6:309-344.

Branco, A. and P. Marrafa, 1999. Long-distance Reflexives and the Binding Square of Opposition. In G. Webelhuth, J. Koening, and A. Kathol (eds.), *Lexical and Constructional Aspects of Linguistics Explanation*. Stanford: CSLI Publications. Chap. 11, 163-177.

Carbonell, J., and R. Brown, 1988. Anaphora Resolution: A Multi-strategy Approach, In *Proceedings, The 12th International Conference on Computational Linguistics (COLING88)*, 96-101.

Chomsky, N., 1980. On Binding. *Linguistic Inquiry*, 11:1-46.

Chomsky, N., 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Dalrymple, M., 1993. *The Syntax of Anaphoric Binding*. Stanford: CSLI Publications.

Garrod, S. and A. Sanford, 1982. The mental representation of discourse in a focussed memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 1:21-41.

Givón, T. 1992. The Grammar of Referential Coherence as Mental Processing Instructions. *Linguistics*, 30:5-55.

Gundel, J., N. Hedberg and R. Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69:274-307.

Gundel, J., 1998. Centering Theory and the Givenness Hierarchy: Towards a Synthesis. In M. Walker, A. Joshi and E. Prince (eds.), pp.183-198.

Guindon, R., 1985. Anaphora Resolution: Short-term Memory and Focusing. In *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics (ACL85),* 218-227.

Lappin, S. and H. Leass, 1994. An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, 20: 535-561.

Löbner, S., 1987. Quantification as a Major Module of Natural Language Semantics. In J. Groenendijk, D. de Jongh and M. Stokhof (eds.), *Studies in DRT and the Theory of Generalized Quantifiers*., Dordrecht: Foris, 53-85.

Mitkov, R., 1997. Factors in Anaphora Resolution: They are not the only Things that Matter. A Case-study based on two Different Approaches. In *Proceedings of the ACL/EACL97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution.* Association for Computational Linguistics.

Mitkov, R., 1998. Robust Pronoun Resolution with Limited Knowledge. In Proceedings, 36th Annual Meeting of the Association of Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL98), 869-875.

Pollard, C. and I. Sag, 1994. *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.

Prince, E., 1981. On the reference of indefinite-*this* NPs. In A. Joshi, B. Webber and I. Sag (eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

Rich, E., and S. LuperFoy. 1988. An Architecture for Anaphora Resolution, In *Proceedings, 2nd Conference on Applied Natural Language Processing*, 18-24.

Walker, M., A. Joshi, and E. Prince. (eds.), 1998a. *Centering In Discourse.* Oxford: Oxford University Press.

Walker, M., A. Joshi, and E. Prince. 1998b. Centering in Naturally Occurring Discourse: An Overview. In M. Walker, A. Joshi and E. Prince (eds.), 1998a, 1-30.

Xue, P., C. Pollard, and I. Sag, 1994. A New Perspective on Chinese *Ziji*. In *Proceedings of the West Coast Conference on Formal Linguistics* (WCCFL'94), Stanford: CSLI Publications.