Cognitive anaphor resolution and the binding principles

António Branco

Abstract

Mainstream cognitive models of nominal anaphor resolution envisage this process as a particular case of search optimisation in the cognitive space of the working memory. In a first step, we uncover the empirical predictions implied by this rationale. Next, we show in which sense these predictions are not satisfied by naturally occurring grammatical classes of anaphors and discuss the impact of this negative result in terms of the empirical grounding of cognitive models of anaphor resolution.

Keywords: reference processing, anaphora, anaphor resolution, binding constraints, binding theory.

1. Introduction

Mainstream cognitive models of nominal anaphor resolution envisage this process as a particular case of search optimization: The search space of working memory with antecedent candidates is “sectioned”, each “section” containing the admissible antecedents for anaphors of different classes. This rationale implies the prediction that there are natural classes of anaphors such that every anaphor in each such class has the same set of admissible antecedents. It further implies that the sets of admissible antecedents induced by these different natural classes of anaphors bear specific relations among them.

The four binding classes of anaphors delimited by means of the grammatical binding principles gathered in the so called “binding theory” are the naturally occurring classes of anaphors such that if the anaphors in the same set occur in whatever syntactic position, they have the same set of admissible antecedents. We verify that the predictions referred to above about the relations between sets of antecedents induced by natural classes of anaphors are not empirically satisfied. We discuss in which sense this negative result casts doubts on the empirical support of the search optimization rationale for the mainstream cognitive accounts of anaphor resolution.

2. Natural classes of anaphors

The search optimization rationale has been the crucial hypothesis explored in cognitive models of anaphor resolution. Differentiation of anaphoric capacity of different anaphors is explained under the assumption that it avoids going through the scanning of the whole working memory in the anaphor resolution process. As speakers refer again, say, to the same person already referred to by *the student with a yellow t-shirt*, the specific anaphoric form they use, e.g. *the student*, *he* or *himself*, depends on the relative position of the representation of the referent of *the student with a yellow t-shirt* in the working memory: Different types of anaphors have thus been assumed to pick referential items from different "sections" of the relevant search space.

For this schema to work, there has to be some feature that discriminates different items in the working memory from one another and induces a partial order over them. This order is typically established according to the attentional prominence that each such item bears. Attentional prominence is assumed to reflect a natural metrics for "distance" in the relevant cognitive search space, with less attentionally prominent items being the ones that take longer to be retrieved.

Skimming through the literature, one finds different proposals concerning the number of sections into which the search space for anaphor 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, Hedberg, & Zacharski 1993), in turn, proposes a schema that may extend the division up to six "sections", depending on the specific language at stake.

3. Predictions

The strong prediction is that anaphors of different types have different, *disjoint* sets of antecedents. This claim can be found, for instance, in (Garrod & Sanford 1982).

Another, weaker but also plausible prediction in this connection is that, if the different sets of admissible antecedents turn out not to be disjoint, they would at least be expected to be *successively included* within each other. If we admit that an anaphor is of a given type such that it 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, for instance, of (Gundel, Hedberg, & Zacharski 1993; Gundel 1998).

The search optimization rationale for anaphor resolution – with the assumed correlation between anaphoric forms and attentional prominence of antecedent candidates – can thus be seen as inducing a delimitation of anaphors into different natural classes. These classes are circumscribed in terms of the antecedents that the corresponding anaphors admit: A given class of anaphors is defined because every anaphor in that class can be resolved against the same set of antecedents.

The point worth stressing here is that this establishes a very interesting and self-contained line of empirical inquiry: If we succeed in isolating different sets 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 we find such natural classes of anaphors, then we are providing a piece of empirical support of paramount importance for the whole conjecture embodied in the search optimization rationale.

4. Failing checking the predictions

A first step towards pursuing this research path is to find a methodological device that allows to categorize items according to their attentional prominence. This involves finding a suitable scale of the attentional prominence of admissible antecedent entities. Besides, we need also objective criteria to decide which item in the scale a given anaphor should be put in correspondence with. The pursuing of these goals has been reported at various places in the literature, cf. among others, (Prince 1981) and (Gundel, Hedberg, & Zacharski 1993).

The scale used to evaluate the attentional status of the cognitive item against which a given anaphor is resolved is typically defined by means of a set of keywords, like “familiar”, “activated”, “evoked”, “uniquely identifiable", "brand new", etc. These keywords come with informal definitions under the form of example sentences and a discussion of some cases to which they may apply. The keywords come also with a hierarchy, where the relative positioning of each keyword in the scale is defined vis a vis the other keywords.

This sort of approach to define a scale of attentional prominence seems to be flawed, in our view, in some crucial aspects: (i) There is not an empirical justification for the number of required keywords, i.e. of the distinct degrees of relevant attentional prominence; (ii) Keywords are defined in such a way that the boundaries between the degrees of prominence they are supposed to delimit are not clear; (iii) Above all, there is no empirically well defined criteria to unequivocally decide which point of the scale is the antecedent of an anaphor in a specific occurrence in correspondence with.

These recurrent shortcomings represent a considerable drawback for the goal of finding empirical support to the search optimization rationale of anaphor resolution. The alternative line of argument we would like to explore in the present article is that overcoming this drawback may involve changing the angle from which the correlation between natural classes of anaphors and search optimization could or should be addressed.

Instead of in the first place looking for objective criteria to identify attentional status of items and then trying to use them to possibly delimit classes of anaphors, we should take into account naturally occurring classes of anaphors – empirically motivated precisely on the basis of differences concerning the classes of their admissible antecedents – and try to clarify the possible cognitive underpinnings of such classes. In particular, one should discuss whether and how such classes may fit into a search optimization rationale for anaphor resolution.

5. Grammatical binding constraints

The most notorious classes of anaphors obtained via grouping of the corresponding sets of admissible antecedents are the so-called binding classes. Each of these classes contains all and only the anaphors that may pick an antecedent from the same set of admissible antecedents. A classical contrast permitting to illustrate the kind of difference at stake is the one between *Peter said John described Tom to* <u>*himself*</u> and

Peter said John described Tom to him: While *himself* have *John* and *Tom* as admissible antecedents but not *Peter*, *him* has Peter as admissible antecedent (and possibly other antecedents introduced in the discourse or the context), but not *John* or *Tom*. Accordingly, *himself* and *him* are said to belong to different binding classes, the former to the class of the so-called short-distance reflexives, the latter to the class of the so-called pronouns.

The members of a given binding class can be intensionally characterized as those anaphors that are ruled by a specific binding constraint, with this constraint expressing an objective criterion to categorize anaphors according to one of the different available binding classes. Such binding constraints capture empirical generalizations and are aimed at delimiting the relative positioning of anaphors and their admissible antecedents in grammatical geometry.

Since their first formulation in (Chomsky 1980, 1981), the definition of binding principles has been the focus of intense research, from which a binding theory of increased empirical adequacy has emerged. From an empirical perspective, binding constraints, or binding principles, 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 modules of grammatical knowledge, usually known under the term of “binding theory”. [1]

Recent developments of (Pollard & Sag 1994), in particular (Xue, Pollard, & Sag 1994; Branco & Marrafa 1999; Branco 2000), indicate that there are four binding constraints. Below, the definition of each principle is illustrated by an example with relevant contrasts: [2]

[1] Vd. (Dopkins & Nordlie 1995) and (van der Lely & Stollwerck 1997) and references cited therein for an overview of psycholinguistic research on binding constraints.

[2] Coindexation marks anaphoric links between anaphors and their tentative antecedent(s); indexes prefixed by ‘*’ mark non-admissible anaphoric links; and ‘{...ant_c...}’ represent tentative antecedents available outside the sentence, in the discourse or in the context. We are using examples of Portuguese, a language with anaphors of each of the four binding classes. Some languages may not have anaphors of every binding type.

Principle A
If a short-distance reflexive is locally o-commanded, it must be locally o-bound.
{*...ant*$_c$*...*} [*O amigo do Lee*$_i$]$_j$ *acha que* [*o vizinho do Max*$_k$]$_l$ *gosta de si próprio*$_{*c/*i/*j/*k/l}$. (Portuguese)
[Lee$_i$'s friend]$_j$ thinks [Max$_k$'s neighbour]$_l$ likes himself$_{*c/*i/*j/*k/l}$.

Principle Z
If a long-distance reflexive is o-commanded, it must be o-bound.
{*...ant*$_c$*...*} [*O amigo do Lee*$_i$]$_j$ *acha que* [*o vizinho do Max*$_k$]$_l$ *gosta dele próprio*$_{*c/*i/j/*k/l}$.
[Lee$_i$'s friend]$_j$ thinks [Max$_k$'s neighbour]$_l$ likes him$_{*c/*i/j/*k}$/himself$_l$.

Principle B
A pronoun must be locally o-free.
{*...ant*$_c$*...*} [*O amigo do Lee*$_i$]$_j$ *acha que* [*o vizinho do Max*$_k$]$_l$ *gosta dele*$_{c/i/j/k/*l}$.
[Lee$_i$'s friend]$_j$ thinks [Max$_k$'s neighbour]$_l$ likes him$_{c/i/j/k/*l}$.

Principle C
A non-pronoun must be o-free.
{*…ant*$_c$*…*} [*O amigo do Lee*$_i$]$_j$ *acha que* [*o vizinho do Max*$_k$]$_l$ *gosta do rapaz*$_{c/i/*j/k/*l}$.
[Lee$_i$'s friend]$_j$ thinks [Max$_k$'s neighbour]$_l$ likes the boy$_{c/i/*j/k/*l}$.

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.[3] For instance, in the

[3] For details, see (Dalrymple 1993).

example *Lee's friend thinks* [*Max's neighbour likes him*] the local domain of *him* is indicated between square brackets.

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 multi-clausal sentence, the upward arguments o-command the successively embedded arguments. For instance, in the example *The girl who said that Peter knows Max thinks Max's neighbour likes him*, we get the following o-command relations: *The girl who said that Peter Max knows Max* < *Max's neighbour* < *him*, and *Peter* < *Max* < *who*.

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.[4] For instance, in the example *Lee's friend thinks Max's neighbour likes himself*, *Lee's friend* (non locally) o-binds *himself*, *Max's neighbour* locally o-binds it, and *Lee* and *Max* does not o-bind it.

Note that, given their conditional definition, Principles Z and A are complied with if the reflexives are in so called *non exempt positions*, that is if they are, respectively, o-commanded and locally o-commanded.

It is now well established in the literature that there is a distinction between constraints for anaphor resolution (excluding tentative antecedents from the set of admissible antecedent candidates) and preferences (making the resolution process to converge on the actual antecedent). Binding constraints are thus to be counted in the set of such constraints, though they are not the only ones.[5]

[4] 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 previous results of ours reported in (Branco 1996), 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).

[5] For details on the distinction between constraints and preferences in anaphor resolution, and their listings, see (Carbonell & Brown 1988; Rich & Luperfoy 1988; Asher & Wada 1988; Lappin & Leass 1994; Mitkov 1997; Branco 2000).

6. Checking failure of predictions

As discussed above, the search optimization rationale for anaphor resolution implies some predictions concerning the relations between the different natural classes of admissible antecedents for anaphors. These classes 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 conform to these predictions. For each of the four binding classes, we delimit the corresponding sets A, B, C and Z of admissible antecedents and then check how they relate to each other.

In order to proceed with this test, first, we have to fix a non exempt position **x** in a generic multi-clausal grammatical structure, like the one used above for the examples illustrating the different binding principles, that can be schematically represented as

$$\{..\text{disc/cont}^{*}..\}..\text{nloc}^{*}..[..\text{noc}^{*}..]..[..\text{loc}^{*}..[..\text{noc}^{*}..]..\mathbf{x}..]_{\text{LocalDom}}..$$

where `nloc*`, `noc*` and `loc*` stand, respectively, for positions of non-local o-commanders, non-o-commanders and local o-commanders. Second, we have to successively instantiate **x** with an anaphor from each different binding class. We will then be able to observe what are the relations among the sets of admissible antecedents of each binding class.

If we assume that **x** is any anaphor complying with principle A, we see that the admissible antecedents of **x** form the set of its local o-commanders, which we can call the set A.

In case **x** is an anaphor complying with principle Z, the set Z of its admissible antecedents is made of its o-commanders.

When **x** is an anaphor ruled by Principle B, the set B of its admissible antecedents contains all the antecedents that are non-local o-commanders of **x**.

Finally, the set C of the admissible antecedents of **x** when this is an anaphor complying with principle C has all the items that are non-o-commanders of **x**.

Given the definitions of the o-command relation, and from a maximally generic point of view, the formal relations between these four sets of admissible antecedents are the following:

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

It is straightforward to see that the admissible antecedents of short-distance reflexives are admissible antecedents of long-distance reflexives ($A \subset Z$); some admissible antecedents of long-distance reflexives are admissible antecedents of pronouns ($Z \cap B \neq \varnothing$); and the admissible antecedents of non-pronouns are admissible antecedents of pronouns ($C \subset B$).

From another perspective, this amounts to say that for a given possible antecedent of an anaphor in position **x**, it is the case that there are always at least two different types of anaphors that can fill **x** and take that antecedent.[6] Or alternatively, for a given anaphor interpreted against a given antecedent, that anaphor can always be replaced at least by another one of a different binding type that can take the same antecedent.

In any case, what is crucial to note for our experiment 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 for anaphor resolution.

7. Conclusions

The search optimization rationale expected to bear on anaphor resolution implies some predictions about the existence of natural classes of anaphors such that, with respect to any position of occurrence, every element in each such class have the same set of admissible antecedents. In particular, it implies that these sets of admissible antecedents exhibit certain relations among them: They are expected either to be disjoint, or at least to be successively included within each other.

Given the current state of the art of the research on anaphora, the four binding classes are the naturally occurring classes of anaphors satisfying the criterion pointed out above: Each binding class contains all

[6] If one considers also exempt syntactic positions, then even reflexives have possible antecedents that may also be antecedents of pronouns and non-pronouns.

and only the anaphors that, for whatever grammatical position, have the same set of admissible antecedents.

The result we argued for in the present paper is that the four sets of admissible antecedents of the four binding classes do not conform to the predictions underlined above: They are neither disjoint nor successively included in each other: While two of them, A and C, are strictly included in the other two, Z and B – with A $\subset$ Z and C $\subset$ B – , the latter are not disjoint neither included in one another. Given that these are objectively determined natural classes of anaphors across natural languages, this result casts serious doubts that the search optimization rationale may provide a clear-cut justification for anaphor resolution and its constraints.

This should not be seen as forcing the inference that cognitively rooted factors (such as attentional prominence associated with recency of mention, just to refer an example) do not play an important role in anaphor resolution, at least as preference factors. This result, if correct, shows that current cognitive models of anaphor resolution, crucially based on the search optimization rationale, make predictions that are infirmed by the very significant empirical generalizations embodied in the definition of binding classes.

References

Asher, N. & H. Wada
1988 A computational account of syntactic, semantic and discourse principles for anaphora resolution. *Journal of Semantics* 6: 309-344.

Branco, A.
1996 Branching split obliqueness at the syntax-semantics interface. In: *Proceedings, 16th International Conference on Computational Linguistics (COLING96)*, 149-156.
2000 *Reference Processing and its Universal Constraints*. Edições Colibri, Lisbon.

Branco, A. & P. Marrafa
1999 Subject-oriented and non subject-oriented long-distance anaphora: an integrated approach. In: *Proceedings, 11th Pacific-Asia Conference on Language, Information and Computation (PACLIC-96)*, 21-31. Seoul.

Carbonell, J. & 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.

1981 *Lectures on Government and Binding*. Foris: Dordrecht.

Dalrymple, M.

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

Dopkins, S. & J. Nordlie

1995 Processes of anaphor resolution. In: R. Lorch and E. O'Brien (eds.), *Sources of Coherence in Reading,* 145-157. Hillsdale: Lawrence Erlbaum.

Garrod, S. & 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 & 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.), 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. & H. Leass

1994 An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20: 535-561.

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.

Pollard, C. & 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 University Press.

Rich, E. & S. LuperFoy

1988 An architecture for anaphora resolution. In: *Proceedings, 2nd Conference on Applied Natural Language Processing*, 18-24.

van der Lely, H. & L. Stollwerck

1997 Binding theory and grammatical specific language impairment in children. *Cognition* 62: 245-290.

Xue, P., C. Pollard & I. Sag

1994 A new perspective on chinese *ziji*. In: *Proceedings of the West Coast Conference on Formal Linguistics (WCCFL'94)*. Stanford: CSLI Publications.