Relative Clause Attachment and Anaphora: A Case for Short Binding

Rodolfo Delmonte
Ca' Garzoni-Moro, San Marco 3417, Università "Ca Foscari", 30124 - VENEZIA
E-mail: delmont@unive.it

Abstract

Relative clause attachment may be triggered by binding requirements imposed by a short anaphor contained within the relative clause itself: in case more than one possible attachment site is available in the previous structure, and the relative clause itself is extraposed, a conflict may arise as to the appropriate s/c-structure which is licenced by grammatical constraints but fails when the binding module tries to satisfy the short anaphora local search for a bindee.

1 Introduction

It is usually the case that anaphoric and pronominal binding take place after the structure building phase has been successfully completed. In this sense, c-structure and f-structure in the LFG framework - or s-structure in the chomskian one - are a prerequisite for the carrying out of binding processes. In addition, they only interact in a feeding relation since binding would not be possibly activated without a complete structure to search, and there is no possible reversal of interaction, from Binding back into s/c-structure level seen that they belong to two separate Modules of the Grammar. As such they contribute to each separate level of representation with separate rules, principles and constraints which need to be satisfied within each Module in order for the structure to be licensed for the following one.

However we show that anaphoric binding requirements may cause the parser to fail because the structure is inadequate. We propose a solution to this conflict by anticipating, for anaphors only the though, the agreement matching operations between binder and bindee and leaving the coindexation to the following module.

In a final section we discuss data from syntactic Treebanks of English – the Penn Treebank – and Italian, the Italian Treebank and the Venice Treebank.

1.1 Positive and Negative Constraints

Anaphoric and Pronominal Binding are usually treated as if they were one single grammatical phenomenon, even though the properties of the linguistic elements involved are quite different, as the subdivision of Binding Principles clearly shows. However, it is a fact, that the grammatical nature of a pronoun - be it an anaphor (short or long one), or a free pronoun - is never taken into account when searching for the antecedent. The anaphoric module of the grammar takes for granted the fact that both the structure associated to the anaphor/pronoun, the grammatical function - at f-structure level in LFG - and the functional features are consistent, coherent and respondent to the Grammaticality constraints stipulated in each grammatical theory. It is the structural level that guarantees consistency, not the Anaphoric/Pronominal Binding Module, which has the only task to add antecedent-pronoun/anaphor indices in the structure, to be used by the semantic modules.

We chose a couple of examples which represent the theoretical query to be solved, given a certain architecture of linguistic theories, which may differ in the way in which they reach a surface representation into syntactic constituents of the input string, but all converge into the need to keep the anaphoric module separate from the structure building process. The examples are in English but may be easily replicated in other languages:

(1) The doctor called in the son of the pretty nurse who hurt herself

(2) The doctor called in the son of the pretty nurse who hurt himself

In the second example we have the extraposition of the relative clause, a phenomenon very common in English but also in Italian and other languages. The related structures theoretically produced, could be the following ones:

(1)a s[np[The doctor],
ibar[called in],
vp[np[the son,

 Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6), pp. 84-89. Universitá di Venezia.

pp[of, np[the pretty nurse,
cp[who, s[pro, ibar[hurt],
vp[sn[herself]]]]]]]]]
(2)a. s[np[The doctor],
ibar[called in],
vp[np[the son,
pp[of, np[the pretty nurse]],
cp[who, s[pro, ibar[hurt],
vp[sn[himself]]]]]]]

If this is the correct input to the Binding Module, it is not the case that 2a. will be generated by a parser of English without special provisions. The structure produced in both cases will be 1a. seen that it is perfectly grammatical, at least before the binding module is applied to the structure and agreement takes place locally, as required by the nature of the short anaphor. It is only at that moment that a failure in the Binding Module warns the parser that something wrong has happened in the previous structure building process. However, as the respective f-structures show, the only output available is the one represented by 2b, which wrongly attaches the relative clause to the closest NP adjacent linearly to the relative pronoun:

2b. s[np[The doctor],
ibar[called in],
vp[np[the son,
pp[of, np[the pretty nurse,
cp[who, s[pro, ibar[hurt],
vp[sn[himself]]]]]]]]]

The reason why the structure is passed to the Binding Module with the wrong attachment is now clear: there is no grammatical constraint that prevents the attachment to take place. The arguments of the governing predicate HURT are correctly expressed and are both coherent and consistent with the information carried out **by the lexical form. At the same time the Syntactic Binding has taken place again correctly by allowing the** empty "pro" in SUBJect position of the relative adjunct to be "syntactically controlled" by the relative pronoun, which is the TOPic binder, in turn syntactically controlled by the governing head noun, the NURSE. There is no violation of agreement, nor of lexical information, nor any other constraint that can be made to apply at this level of analysis in order to tell the parser that a new structure has to be produced.

2 Parsing Strategies and Preferences

In order for a parser to achieve psychological reality it should satisfy requirements coming simultaneously from three different fields/areas: psycholinguistic plausibility, computational efficiency in implementation, grammatical constraints. Principles underlying the parser architectures should not belong exclusively to one or the other field disregarding issues which might explain the human processor behaviour. Principles are bestowed psychological reality in performance whenever they may be safely tested, on a statistically relevant sample of individuals. So we annect a lot of importance to the fact that the parser actually behaves like what is expected with human processors. In this case and only in this case we say that the principles are predictive and that the parser we implemented is actually relevant for a theory of parsing.

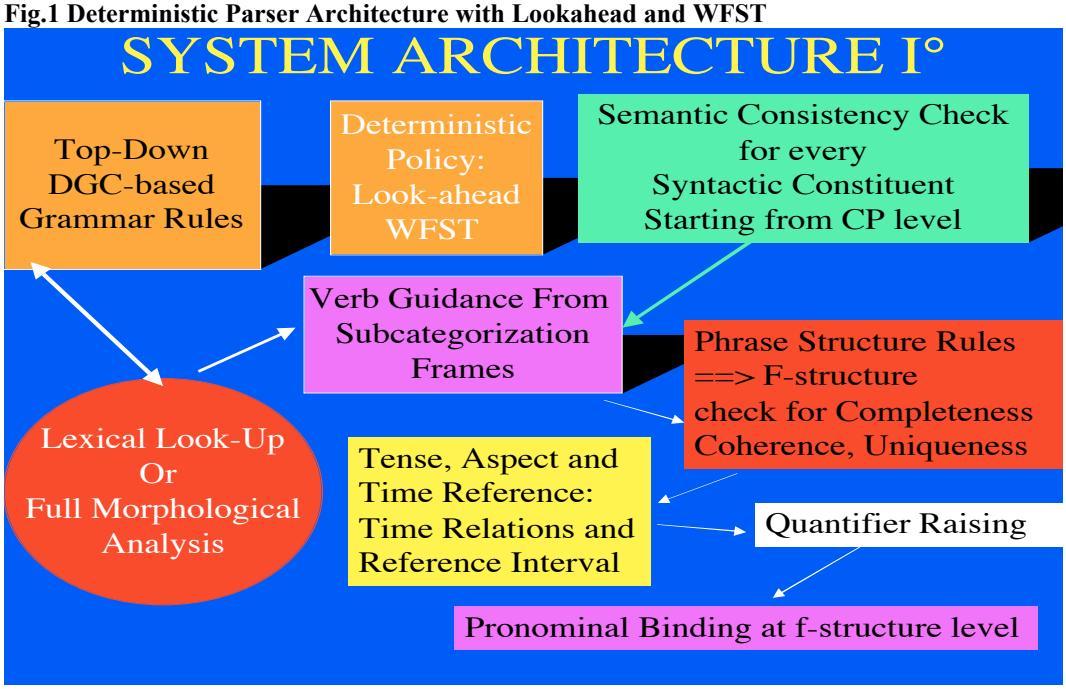
Among contemporary syntactic parsing theories, the garden-path theory of sentence comprehension proposed by Frazier(1987a, b), Clifton & Ferreira (1989) among others, is the one that most closely represents our point of view. It works on the basis of a serial syntactic analyser, which is top-down, depth-first - i.e. it works on a single analysis hypothesis, as opposed to other theories which take all possible syntactic analysis in parallel and feed them to the semantic processor.

Differently from what is asserted by global or full paths approaches (see Schubert, 1984), we believe that decisions on structural ambiguity should be reached as soon as possible rather than deferred to a later level of representation. In particular, Schubert assumes "...a full paths approach in which not only complete phrases but also all incomplete phrases are fully integrated into (overlaid) parse trees dominating all of the text seen so far. Thus features and partial logical translations can be propagated and checked for consistency as early as possible, and alternatives chosen or discarded on the basis of all of the available information(ibid., 249)." And further on in the same paper, he proposes a system of numerical 'potentials' as a way of implementing preference trade-offs. " These potentials (or levels of activation) are assigned to nodes as a function of their syntactic/semantic/pragmatic structure and the preferred structures are those which lead to a globally high potential. Other important approaches are represented by Hindle et al., 1993, who attempt to solve the problem of attachment ambiguity in statistical terms. The important contribution they made, which was not possible in the '80s, is constituted by the data on attachment typologies derived from syntactically annotated corpora.

Our parser copes with ambiguity while at the same time allowing for psychological coherence. Parser architecture is presented in Fig.1 below. The structures produced by the parser take only different processing time to allow for backtracking to take place within the main parser body: but then the right attachment is achieved and the complete structure is produced with the right binding.

We implemented two simple enough mechanisms in order to cope with the problem of nondeterminism and backtracking. At bootstrapping we have a preparsing phase where we do lexical lookup and we look for morphological information: at this level of analysis of all input tokenized words, we create a stack of pairs input wordform - set of preterminal categories, where preterminal categories are a proper subset of all lexical categories which are actually contained in our lexicon. The idea is simply to prevent attempting the construction of a major constituent unless the first entry symbol is well qualified. When consuming any input wordform, we remove the corresponding pair on top of stack.

Fig.1 Deterministic Parser Architecture with Lookahead and WFST



In order to cope with the problem of recoverability of already built parses we built a more subtle mechanism that relies on Kay's basic ideas when conceiving his Chart(see Kay, 1980). Differently from Kay, however, we are only interested in a highly restricted topdown depthfirst parser which is optimized so as to incorporate all linguistically motivated predictable moves. An already parsed RC is deposited in a table lookup accessible from higher levels of analysis and consumed if needed. To implement this mechanism in our DCG parser, we assert the contents of the RC structure in a table lookup storage which is then accessed whenever there is an attempt on the part of the parser to build up a RC. In order to match the input string with the content of the store phrase, we implemented a WellFormed Substring Table(WFST) as suggested by Kay(1980).

Now consider the way in which a WFST copes with the problem of parsing ambiguous structure in his chart. It builds up a table of well-formed substrings or terms which are partial constituents indexed by a locus, a number corresponding to their starting position in the sentence and a length, which corresponds to the number of terminal symbols represented in a term. For our purposes, two terms are equivalent in case they have the same locus and the same length. In this way, the parser would consume each word in the input string against the stored term, rather than against a newly built constituent. In fact, this would fit and suit completely the requirement of the parsing process which rather than looking for lexical information associated to each word in the input string, only needs to consume the input words against a preparsed well-formed syntactic constituent.

However, in order for a parser to show coherent psychological behaviour it should show “garden path” effects while simulating a condition of failure to parser at propositional level (see Pritchett). Full paths parsers, and in general all bottom-up chart-like parsers will not show any of the garden-paths effects simply because failure is prevented from taking place by the fact that all possible parses are always available and can be retrieved by the parser itself. The question that is posed by our two examples will not however be covered by a full-paths parser seen that there is no principled reason in the grammar to prefer one structure over the other. Failure only takes place in the pronominal binding module which is usually a separate module of the parser.

3 Short Anaphora

The parser we use has shown the effect of "garden path", in that it has gone into a loop with the unwanted result of "freezing" the computer, due to data overflow. In other words, as soon as the Binding Module tries to process the f-structure received as input, seen that short anaphora requires binding to take place within a local domain, f-command - the corresponding c-command in functional terms, applied to grammatical functions and a graph structure - will impose the same level of containment for both the pronoun and the antecedent. And seen that the only antecedent available is the empty SUBJect which has functional features inherited by means of syntactic control from the governing relative pronoun, the agreement match is attempted, and a failure ensues systematically.

As a result of a failure at the Binding Level, a call to the structural level is issued which attempts to build the structure another time. But seen that no failure has taken place at this level of analysis, the result will be the same as the previous one. And this process will go on indefinitely, seen that the two modules obey different Principles and satisfy them separately.

We will now put forward a theoretical proposal regarding exclusively short anaphors, thus disregarding long anaphors and reciprocals in particular or "proprio" in Italian, which call for a different treatment. The proposal we will make is very simple:

"short anaphora must be checked for agreement with their available binder already at the level of satisfaction of grammatical principles, before the structure is licensed"

This requirement is not introduced by the need to improve on the implementation side of the parser, but responds to theoretical principles inherent in the formulation of the Binding Principles. Short anaphora not only obey positive constraints, as opposed to the other pronominals, they also carry a locality requirement which is equivalent to the same domain in which Grammaticality Principles apply, such as the ones expressed in LFG - Uniqueness, Completeness, Consistency. At each "propositional" level, corresponding to a simple f-structure and roughly to a Complete Functional Complex in GB terms(see Chomksy 1986, 169), all arguments of the governing predicate must be checked for completeness - they must all be present at functional level, even if they may be lexically empty; they must be coherent, only those included in the corresponding lexical form must be present; each functional attribute must be assigned to a unique functional value. And in our case no violation is detectable seen that the attributes belonging to the empty "pro" SUBJect are unique even though they are not appropriate to bind the short anaphor OBJect of the same predicate HURT. However there is no indication in the grammar that they should be checked for agreement at this level of analysis.

By anticipating the working of the Binding Module, we assume that Short Anaphors belong partly to the Grammar level and partly to the Binding level: they belong to the grammar level since they require and can to be licensed at sentence or propositional level withouth their f-features being in agreement with their antecedent and binder. Besides, they belong to the binding level where agreement takes place and coindexation follows, in case of success.

As to cases in which the anaphor is contained within a NP in SUBJect position of a sentential complement, the search for the antecedent is suspended not being available locally and no agreement match can be performed. This will not apply to anaphors contained within the NP of the OBJect seen that the antecedent is available.

A failure in the Anaphoric Module will simply cause the Parser to backtrack but the structure produced will not change seen that the failure has taken place in a separate module. Of course, the alternative is using a single unification mechanism that takes context-free rules with all possible alternatives, builds a tentative structure than unifies functional features, and in case of failure tries another possible structure. However this perspective is not only computationally inefficient, it is basically psychologically unfeasible: there will be no principled reason to tell Garden Path sentences apart from the rest seen that all sentences can be adjusted within the parser, sooner or later. Also processing time is not controllable seen that the parser will produce all possible structures anyway and there is no way to control the unification mechanism in a principled manner. On the contrary, in a parser like ours, the order of the rules is controlled strictly, and also the way to produce backtracking is controlled, seen that the parser has a lookahead mechanism that tells the parser which rule to access or not at a given choice point.

Going back to our couple of examples of the Extraposed Relative Clause containing a Short Anaphor, the question would be to prevent Failure since we do not want Constituent Structure Building to be dependent upon the Binding of the Short Anaphor. The only way out of this predicament is that of anticipating in Sentence Grammar some of the Agreement Checking Operations as proposed above. So the Parser would be able to backtrack while in the Grammar and to produce the attachment of the Relative Clause at the right place, in the higher NP headed by the masculine N, “the son”. The important result would be that of maintaining the integrity of Syntax as a separate Module which is responsible in “toto” of the processing of constituent structures. The remaining Modules of the Grammar would be fully consistent and would use the information made available in a feeding relation, so that interpretation will follow swiftly.

To integrate this suggestion coming from Implementation problems, into the theoretical Framework of LFG or other similar theories we simply need to integrate GRAMMATICALITY PRINCIPLES as they have been stipulated so far, to be consisting of:

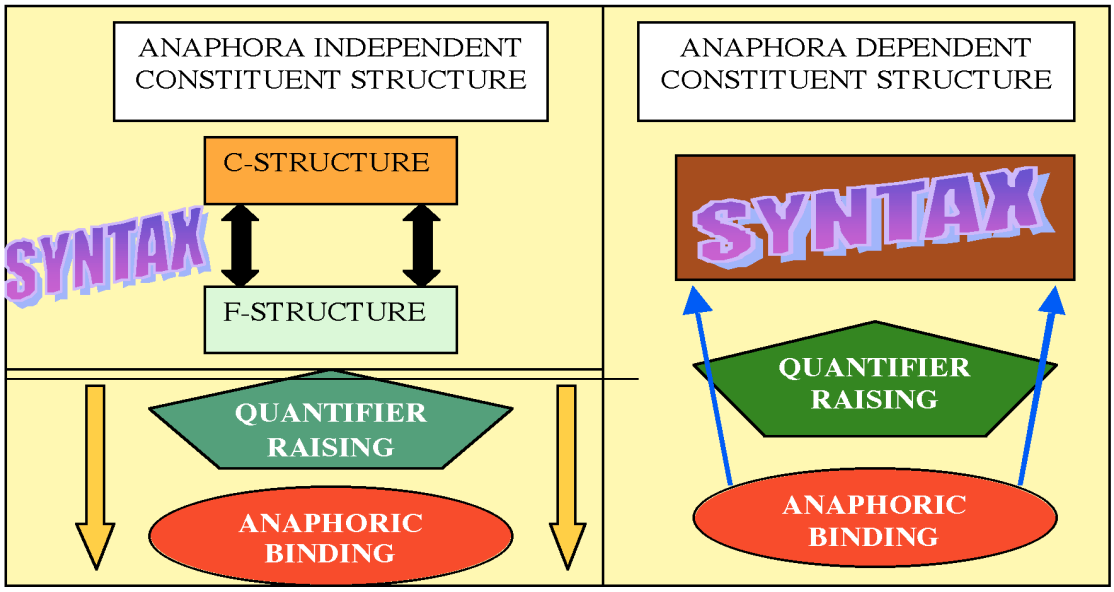
- UNIQUENESS; COHERENCE; COMPLETENESS

with the additional restriction:

- BOUND ANAPHORA AGREEMENT

i.e. short anaphors should be checked before leaving sentence grammar, for agreement with their antecedents iff available in their Minimal Nucleus. In particular, seen that in our framework Quantifier Raising is performed before Anaphoric Binding and will produce new arcs in the graph to represent the scope of quantifiers, this will also undergo failure in order to try a new analysis. This is both time-consuming and unrealistic. A simpler way to solve this problem is to introduce Short Binding as has been defined above. In this way we split Bound Anaphors and make them obey the same principles of Sentence Grammar to which they belong in all respect. In Fig.2 below we show how anaphoric binding and grammatical principles interact. In Fig.2b Anaphoric Binding interacts with Syntax thus causing a failure to take place which cannot be recovered seen that there are other intervening parsing modules. In Fig.2a, on the contrary we postulate the separation between the output of the syntax to be fully autonomous from QR and AB, thus resulting in a more efficient and psychologically viable simulation.

Fig.2a Anaphora Independent Syntactic Parsing **Fig.2b Anaphora Dependent Syntactic Parsing**



4 Experimental Results from Treebanks

We decided to look at corpus data derived from available treebanks in order to ascertain whether the phonemon we are modeling is actually present in real texts. We also wanted to verify whether the RC extraposition was subject to variation from one language to another. We searched in the available treebanks, PennTreebank for English with 1,000,000 tokens, and the Treebank of Italian we are currently working in for syntactic constituenty XML annotation as well as the Venice Treebank made up of approximately the same number of tokens for a total of 300,000 tokens.

We considered only relative clause with morphologically expressed complementizer, thus disregarding all reduced relative clauses. As to the distinction between extraposed vs. non-extraposed we simply looked at the number of brackets – only one - intervening between the constituent label introducing the relative clause in PennTreebank, which is the following (SBAR (WH, and none in the VeniceTreebank. For all remaining cases we counted an extraposed RC.

We tabulated the results in the Table 1. below where we see that Italian is a language much richer on Relative Clauses than American English. In particular the amount of relative clauses in the Italian Venice Treebank is 3 times that of the PT. Yet more interesting seems the ratio of Head Adjacent vs. Non Head

Adjacent RCs: we see that here, whereas Italian has 1 potentially ambiguous RC every 4 RCs, PT has 1 every 6. We can thus conclude that Italian is much more ambiguous to be parsed than English as far as relative clause attachment is concerned.

Table 1. Treebank Derived Structural Relations for Relative Clauses

	Total No. Tokens	Total No. Sentences	Total No. Rel.Cls.	Head Adjacent	Non Head Adjacent
PENN Treebank	**1,000,000**	**44808**	**11559**	**8906 = 77.05%**	**2653 = 22.9%**
SUSANNE Corpus	**130,000**	**7912**	**1380**	**1089 = 78.9%**	**291 = 21.1%**
VENICE Treebank	**300,000**	**11108**	**5155**	**3867 = 75%**	**1288 = 25%**

However, the most interesting fact is constituted by the proportion of relative clauses in relation to the total number of sentences: Generic American English in the Susanne Corpus, counts 1 relative clause every 5/6 sentences; Specialized American English in the PT, goes up to 1 relative clause every 4 sentences. Italian raises the proportion to one relative clause every 2 sentences. Data reported by J.Fodor are in favour of a Head Adjacent use of RC in English, being a language governed by a phonologically related strategy of Minimal Attachment which tends to prevent RC Extraposition. This state of affairs would have RC production in English more restricted than in languages like Italian, which allow for multiple syntactic binders, both adjacent and non-adjacent ones. Data reported in Table 2. seem to support this hypothesis.

Table 2. Treebank Derived Structural Relations for Relative Clauses

	Total No. Tokens	Total No. Sentences	Total No. Rel.Cls.	Complex Relative Clauses	Ratio Rel.Cls. / Sentences	Ratio Rel.Cls /Tot.Tokens
PENN Treebank	**1,000,000**	**44808**	**11559**	**2724**	**25.8%**	**1.16%**
SUSANNE Corpus	**150,000**	**7912**	**1380**	**106**	**17.44%**	**0.92%**
VENICE Treebank	**300,000**	**11108**	**5155**	**-**	**46.40%**	**1.72%**

References

Clifton C., & F. Ferreira(1989), Ambiguity in Context, in G.Altman(ed), Language and Cognitive Processes, op.cit., 77-104.

Delmonte R., D.Bianchi(1991), Binding Pronominals with an LFG Parser, Proceeding of the Second International Workshop on Parsing Technologies, Cancun(Messico), ACL 1991, pp. 59-72.

Delmonte R.(2000), Generating and Parsing Clitics with GETARUN, Proc. CLIN'99, Utrech, pp.13-27.

Delmonte R.(2000),(to appear), Parsing Preferences And Linguistic Strategies, Proc.Workshop Communicating Agents, IKP, Bonn, pp.15.

Delmonte R.(2000), Parsing with GETARUN, Proc.TALN2000, 7° confèrence annuel sur le TALN,Lausanne, pp.133-146.

Fodor J.(2002), Psycholinguistics cannot escape prosody, Invited Talk, SpeechProsody2002, Aix-en-Provence.

Frazier L.(1987a), Sentence processing, in M.Coltheart(ed), Attention and Performance XII, Hillsdale, N.J., Lawrence Elbaum.

D.Hindle & M.Roth(1993), Structural Ambiguity and Lexical Relations, Computational Linguistics 19, 1, 103-120.

Schubert L.K.(1984), On Parsing Preferences, Proc. of COLING, 247-250.

Kay Martin(1980), Algorithm Schemata and Data Structures in Syntactic Processing, CSL-80-12, Xerox Corporation, Palo Alto Research Center.

Pritchett B.L.(1992), Grammatical Competence and Parsing Performance, The University of Chicago Press, Chicago.