Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm

Roland Stuckardt*
Knowbotic Systems GmbH & Co. KG

Syntactic coindexing restrictions are by now known to be of central importance to practical anaphor resolution approaches. Since, in particular due to structural ambiguity, the assumption of the availability of a unique syntactic reading proves to be unrealistic, robust anaphor resolution relies on techniques to overcome this deficiency.

This paper describes the ROSANA approach, which generalizes the verification of coindexing restrictions in order to make it applicable to the deficient syntactic descriptions that are provided by a robust state-of-the-art parser. By a formal evaluation on two corpora that differ with respect to text genre and domain, it is shown that ROSANA achieves high-quality robust coreference resolution. Moreover, by an in-depth analysis, it is proven that the robust implementation of syntactic disjoint reference is nearly optimal. The study reveals that, compared with approaches that rely on shallow preprocessing, the largely nonheuristic disjoint reference algorithmization opens up the possibility for a slight improvement. Furthermore, it is shown that more significant gains are to be expected elsewhere, particularly from a text-genre-specific choice of preference strategies.

The performance study of the ROSANA system crucially rests on an enhanced evaluation methodology for coreference resolution systems, the development of which constitutes the second major contribution of the paper. As a supplement to the model-theoretic scoring scheme that was developed for the Message Understanding Conference (MUC) evaluations, additional evaluation measures are defined that, on one hand, support the developer of anaphor resolution systems, and, on the other hand, shed light on application aspects of pronoun interpretation.

1. Introduction

The interpretation of anaphoric expressions is known to be a difficult problem. In principle, a variety of constraints and preference heuristics, including factors that rely on semantic, pragmatic, and world knowledge, contribute to this task (Carbonell and Brown 1988). Robust, operational approaches to anaphor resolution on unrestricted discourse, however, are confined to strategies exploiting globally available evidence like morphosyntactic, syntactic, and surface information.

Beginning with the pioneering work of Hobbs (1978), many practical approaches rely on the availability of syntactic surface structure by employing coindexing restrictions, salience criteria, and parallelism heuristics (e.g., Lappin and Leass 1994). However, even the assumption of the availability of a unique syntactic description is unrealistic since, in general, parsing involves the solution of difficult problems like at-

* Daimlerstraße 32, D-60314 Frankfurt am Main. E-mail: roland@stuckardt.de

tachment ambiguities, role uncertainty, and the instantiation of empty categories. Based on this observation, Kennedy and Boguraev (1996) have suggested an adaptation of Lappin and Leass's approach to the shallow analysis frontend of English Constraint Grammar (Karlsson et al. 1995), which provides a part-of-speech tagging comprising an assignment of syntactic function but no constituent structure. This information deficiency is partly overcome by the application of a regular filter that heuristically reconstructs parts of the constituent structure. An alternative solution, which is based on the possibly partial but potentially more comprehensive and reliable output of a conventional parser, has been suggested in Stuckardt (1997).

In the present paper, an approach to robust anaphor resolution is developed that enhances the latter work. The coreference resolution algorithm ROSANA[1] is developed, the core of which consists of a set of rule patterns by means of which the verification of disjoint reference rules is generalized in order to make it applicable to deficient (fragmentary) syntactic descriptions. Based on this algorithm, the ROSANA system, which works on the partial syntactic descriptions generated by the robust FDG (Functional Dependency Grammar of English) parser of Järvinen and Tapanainen (1997), is implemented. By a formal evaluation on two text corpora that differ with respect to genre and domain, it is proven that ROSANA achieves robust (truly operational) high-quality coreference resolution on unrestricted texts. An in-depth analysis shows that the robust implementation of syntactic disjoint reference is nearly optimal. Compared with approaches that rely on a combination of shallow preprocessing and heuristic syntactic disjoint reference, the largely nonheuristic disjoint reference algorithmization employed by ROSANA opens up the possibility for a slight improvement.

The performance study of the ROSANA system crucially rests on an enhancement of the evaluation methodology for coreference resolution systems, the development of which constitutes the second major contribution of the paper. As a supplement to the coreference class scoring scheme that was developed for the CO-task evaluation of the Message Understanding Conferences (Vilain et al. 1996), two additional evaluation disciplines are defined that, on one hand, aim at supporting the *developer* of anaphor resolution systems, and, on the other hand, shed light on *application aspects* of pronoun interpretation. The evaluation of ROSANA according to the refined scoring scheme gives evidence that the interpretation quality may be improved by a genre-specific choice of the preference factors and their relative weights. This demonstrates the usefulness of enhancing the evaluation methodology for coreference resolution systems.

The paper is organized as follows. In Section 2, the robustness issue of natural language processing is briefly discussed at a general level, and two models of robust anaphor resolution are introduced. In Section 3, by deriving a set of disjoint reference rule patterns for fragmentary syntax, the core component of a robust, operational anaphor resolution algorithm is developed. In Section 4, the ROSANA algorithm is designed, and an implementation, the ROSANA system, is described. In Section 5, an enhanced set of evaluation disciplines for coreference resolution systems is advocated for, and the respective formal measures are defined. In Section 6, the evaluation results of ROSANA are discussed. Finally, in section 7, ROSANA is compared with other approaches to anaphor resolution and, in particular, robust syntactic disjoint reference.

1 ROSANA = **Ro**bust **S**yntax-Based Interpretation of **Ana**phoric Expressions

2. Anaphor Resolution and the Robustness Issue

2.1 Robustness in Natural Language Processing

In natural language processing in general, the robustness issue comprises the ability of a software system to cope with input that gives rise to **deficient descriptions** at some descriptive layer.[2] More or less implicit is the assumption that the system exhibits some kind of *monotonic behavior*: the less deficient the description, the higher the quality of the output (Menzel 1995).

Following Menzel further, this intuitive characterization may be refined. The processing should exhibit *autonomy* in the sense that complete failures at one stage of analysis should not cause complete failures at other stages of analysis or even a failure of the overall processing. Moreover, the processing model should ideally employ some kind of *interaction* between different stages of analysis: deficiency at one stage of analysis should be compensated by evidence gained at other stages.

2.2 Two Models of Robust Anaphor Resolution

In light of the above characterization, the robustness requirement for the anaphor resolution task may be rendered more precisely. In the aforementioned operational approaches, a *sequential processing model* is followed according to which anaphor resolution is performed by referring to the result of an *already completed* syntactic analysis. Under this architecture, evidence for structural disambiguation that may be contributed by strong expectations at the referential layer is not taken into account (Stuckardt 1996). In terms of the general goals of robust processing, since there is no interaction, this implies that the robustness requirement only shows up in the form of the monotonicity and autonomy demands: the anaphor resolution module has to cope with deficient or shallow syntactic information. Besides the trivial way of achieving this kind of robustness by simply not exploiting deficient syntactic descriptions, the following two models may be employed:

- **the shallow description model:** by employing heuristic rules to partially reconstruct the syntactic description, the anaphor resolution strategies are adapted to *shallow input data* that are never defective;[3]
- **the deficient description model:** by extending anaphor resolution strategies to work on a possibly ambiguous or incomplete description, *syntactic evidence is exploited as far as available*.

In contrast to the approach of Kennedy and Boguraev (1996), which follows the shallow description model, the ROSANA algorithm is based on the deficient description model. In principle, even a "complete" algorithm that establishes the conceptually superior degree of robustness by means of interaction between structural disambiguation and anaphor resolution is feasible (Stuckardt 1997). As will become evident, however, the technically less complex sequential strategy employed by ROSANA already yields high-quality results and does not leave much room for further improvement.

2 The deficiency may result either because the input itself is deficient or from shortcomings of the processing resources (e.g., lexicon, grammar/parser, or semantic/pragmatic disambiguation).

3 Here, the monotonicity demand of intuitive robustness virtually vanishes, since there is no longer a syntactic input prone to deficiency.

2.3 Fragmentary Syntax

The main phenomena that give rise to structural ambiguity of syntactic descriptions are *uncertainty of syntactic function* (involving subject and direct object) and *attachment ambiguities* of prepositional phrases, relative clauses, and adverbial clauses. In example (1),

(1) Peter observed the man with the telescope.

depending on the availability of disambiguating information, it may be uncertain whether the underlined prepositional phrase *with the telescope* should be interpreted adverbially or attributively. From the configurational perspective, these ambiguities give rise to **fragmentary syntactic descriptions** that consist of several tree-shaped components. With the exception of the topmost tree fragment, all components correspond to a constituent of type PP, S, or NP whose attachment or role assignment failed.

In addition, *cases in which no reading exists* give rise to fragmentary syntactic descriptions comprising the constituents whose combination failed due to constraint violation.

2.4 Fragmentary Syntax and Anaphor Resolution

Among the anaphor resolution strategies potentially affected by fragmentary syntax are both heuristics and constraints. Preference criteria like salience factors and syntactic parallelism are not affected by all types of syntactic defects. Moreover, there are many heuristics that do not rely on syntactic function or structure. Structural coindexing constraints, however, may lose evidence in all the above cases of fragmentary syntax. Since they are known to be of central importance to the antecedent-filtering phase of operational anaphor resolution approaches, the subsequent discussion focuses on the impact of deficient surface structure description on disjoint reference restrictions.

By referring to Chomsky's Government and Binding (GB) Theory, the core of the syntactic coindexing restrictions may be stated as follows (Chomsky 1981):[4]

Definition
Binding principles A, B, and C:

(A) A reflexive or reciprocal is bound in its binding category.

(B) A pronoun is free (i.e., not bound) in its binding category.

(C) A referring expression[5] is free in any domain.

where **binding category** denotes the next surface-structural dominator containing some kind of subject, and **binding** is defined as *coindexed and c-commanding*:

4 Various theoretical models that cover disjoint reference phenomena have been stated. Since the disjoint reference conditions are descriptive principles of grammar, the choice of the theoretical model is, in this sense, arbitrary. In the subsequent discussion, the comprehensive and widely known GB Theory is referred to.

5 For example, common nouns and names.

Definition
Surface structure node X *c-commands* node Y if and only if the next branching node that dominates X also dominates Y and it is not the case that X dominates Y, Y dominates X, or X = Y.

The following examples illustrate the scope of the binding principles:

(2) a. The barber$_i$ is shaving himself$_i$/$*$him$_i$.

b. The client$_i$ wants that the barber$_j$ shaves $*$himself$_i$/him$_i$.

c. $*$The client$_i$ wants that the barber$_j$ shaves the client$_i$.

In sentence (2a), whereas the reflexive *himself* is required to be coindexed with the local subject *the barber* (binding principle A), coindexing the pronoun *him* with the subject is ruled out (binding principle B) because, otherwise, the pronoun would be locally bound in its binding category. Sentence (2b) illustrates the case of nonlocal binding (here: outside the embedded sentence), which is admissible only in the case of the nonreflexive pronoun. As illustrated by sentence (2c) and modeled by binding principle C, referring expressions (e.g., common nouns and names) are not even allowed to be bound nonlocally.

A further structural well-formedness condition, commonly called the **i-within-i condition**, aims at ruling out certain instances of referential circularity, that is, coindexings matching the pattern $[\alpha \ldots [\beta \ldots]_i]_i$ (Chomsky 1981, page 212). It is motivated by cases like (3):

(3) $*$Mary knows [the owner of his$_i$ boat]$_i$.

The following example illustrates that syntactic fragmentation may interfere with the application of syntactic disjoint reference conditions:

(4) Peter observed the owner of the telescope with it.

In (4), the i-within-i condition possibly applies:[6] the coindexing of *telescope* and *it* is admissible only if the PP containing *it* is not interpreted as an attribute to *telescope*—otherwise, in violation of the i-within-i condition, the pronoun would be contained in the NP of the tentative antecedent. Hence, if the PP attachment ambiguity has not

6 Since the (maximal projection of the) NP with head *telescope* dominates the NP of *it*, the former NP does not c-command the latter. Hence, coindexing the two NPs does not induce a relation of (local) binding, which implies that binding principle B does not apply in this case.

If one assumes the applicability of the i-within-i condition defined as above, however, there are two classes of cases that do not seem to be appropriately distinguished:

(i) Peter observed the owner of [the telescope near its$_i$ factory]$_i$.

(ii) $*$Peter destroyed [a picture of its$_i$ frame]$_i$.

Whereas in (i), where the possessive occurs in an adjunct phrase of the *telescope* NP, coindexing may be judged admissible, in (ii), where the possessive occurs in a complement phrase, coindexing seems to be inadmissible. If, however, the scope of the i-within-i condition were restricted to the complement cases (ii), then, since binding principle B does not apply, case (4) would remain unaccounted for. Whereas, in theory, it is widely agreed that the original definition of the i-within-i condition may be somewhat too strong (e.g., Chomsky 1981, page 212), with respect to the practical task of robust anaphor resolution, as the formal evaluations below will demonstrate, the original i-within-i condition is sufficient. In corpora, cases like (i) seem to be exceptional.

been resolved prior to anaphor resolution, the fragmentary syntactic description does not contribute the configurational evidence necessary for definitely confirming the antecedent candidate *telescope*.

3. Checking Binding Constraints on Fragmentary Syntax

3.1 Basic Observations

The first step toward the verification of binding constraints on fragmentary syntax is suggested by the following observation: *if both the anaphor and the antecedent candidate are contained in the same connected component of the fragmentary syntactic description, no (direct) binding-theoretic evidence is lost*. In this case, it will be possible to verify the binding restrictions of anaphor and antecedent in a nonheuristic manner, since the necessary positive (→ binding principle A) and negative (→ binding principles B, C) syntactic-configurational evidence is entirely available.[7] However, even in the disadvantageous case in which the anaphor and the antecedent candidate occur in different surface structure fragments, a closer look at the fragments may reveal additional information.

3.2 An Example

The following example illustrates a typical case:[8]

(5) Der Mann hat den Präsidenten besucht, der ihn von sich überzeugte.
the man has the president visited who him of himself convinced
'The man visited the president who convinced him of himself.'

Because of the intervening past participle, the relative clause may be interpreted as an attribute to either *Mann* or *Präsidenten*. Hence, syntactic ambiguity arises, yielding a surface structure description that consists of the following two fragments:[9]

(S Mann
(VP Präsidenten))
(S der
(VP ihn
(VP (PP sich))))

In addition, it is known that the second fragment is embedded in the first. There are three pronominal anaphors to be resolved: the reflexive pronoun *sich* of binding-theoretic type A, the nonreflexive pronoun *ihn* of type B, and the relative pronoun *der* of type B.

Regarding the reflexive pronoun *sich*, it can be shown that binding-theoretic evidence is completely available. Clearly, this holds with respect to the candidates *der* and *ihn*, which are contained in the same surface structure fragment. However, even regarding the two candidates *Mann* and *Präsidenten* that occur in the other fragment, there is no loss of evidence: since the reflexive pronoun is of binding-theoretic type A, and the fragment in which it occurs contains its binding category (the S node of the

7 However, this statement applies solely to the direct comparison of the involved occurrences, since in case of further, transitive coindexings, negative evidence stemming from decision interdependency may get lost (cf. Section 4.1).
8 The example is given in German because the structural ambiguity emerges more strikingly.
9 For the sake of readability, parts of the constituent structure are omitted.

relative clause), according to binding principle A both candidates may be definitively *ruled out*.

Similar observations can be made regarding the pronouns *ihn* and *der*, for which binding principle B applies: the two candidates *Mann* and *Präsidenten* are recognized as configurationally *admissible*. In this case, besides the binding category condition, it is decisive that their fragment is known to be *embedded* in the antecedent's fragments.[10]

3.3 Rule Patterns

In the subsequent discussion, pairs of anaphors α and antecedent candidates γ are considered that occur in different surface syntactic fragments. The goal is to determine whether coindexing α and γ (as in the case of actually choosing γ as the antecedent of α) complies with the above-stated binding-theoretic conditions. Since, according to the definition of the binding principles, no asymmetric distinction between anaphor and candidate is drawn, the disjoint reference requirements of both α and γ must be taken into account.

By an abstraction over cases like the ones discussed in Section 3.2, a set of **rule patterns** can be designed by means of which the verification of syntactic disjoint reference is generalized in order to make it applicable to fragmentary syntactic descriptions (cf. Figure 1).[11] As illustrated by example (5), there are two classes of patterns.[12] One class (five patterns, labeled "$*$") matches cases in which, according to the binding principles, coindexing the anaphor α and the antecedent candidate γ is *ruled out*; the other class (three patterns, labeled "$\surd$") applies in certain cases where no binding principle is violated and coindexing is therefore *admissible*. By the binding principles, conditions regarding, on one hand, the *presence or absence* of a c-command relation, and, on the other hand, the *locality or nonlocality* of this relation, are stated. The rule patterns are designed to match fragmentary cases in which at least one condition of either anaphor or candidate is violated ("$*$" patterns), or cases in which all conditions of anaphor and candidate are satisfied ("$\surd$" patterns). Figure 2 explicates the specific conditions the different patterns aim at. Three patterns apply in certain cases of binding principle A violation ([E2]: missing locality; [E4]: missing c-command relation; [F2]: either missing locality or missing c-command relation). Another two patterns cover instances of binding principle C violation ([E3a], [E3b]: c-command relation). Three other patterns match cases of binding principle B satisfaction ([F1], [E1a], [E1b]: nonlocality).

This collection of rules may be supplemented with further patterns employing more sophisticated conditions regarding the fragments to be matched.[13] The actual choice of rule patterns, however, should depend on the parser that is used. As will

10 It is evident that there are cases in which the latter condition does not hold and the coindexing would violate binding principle C.

11 The following notational conventions are used: round brackets delimit constituents; square brackets emphasize fragment boundaries; $bc(X)$ denotes the binding category of surface structure node X; $bn(X)$ denotes the branching node dominating X according to the c-command definition; the subscript of $X_{type\,Y}$ denotes that the binding-theoretic class of the occurrence contributed by X is $Y \in \{A, B, C\}$; for example, $P_{type\,B}$ is a pronoun. $\surd/*$ indicate the prediction of the particular pattern, that is, whether, in structural configurations matching the pattern, coindexing is admissible or ruled out.

12 Example (5) illustrates an instance of syntactic fragmentation that is due to structural ambiguity. The rule patterns, however, are general in the sense that they also cover cases of fragmentary syntactic description that are induced by parsing constraint violation (cf. Section 2.3).

13 For example, patterns may be added that match instances of *binding principle B violation*, that is, cases in which one occurrence of type B is *locally c-commanded* by the other occurrence. To recognize such cases, *two* conditions must be verified, one requiring that one occurrence c-command the other (of type B) regardless of the attachment choice, the other requiring that the fragment of the latter occurrence *not* contain the occurrence's binding category. Similar conditions may be employed for recognizing instances of *binding principle A satisfaction*.

[F1]	$\surd$	$\{ \ldots F_i = [\ldots bc(\gamma)(\ldots \gamma_{type\,B} \ldots) \ldots] , \ldots , F_j = [\ldots bc(\alpha)(\ldots \alpha_{type\,B} \ldots) \ldots] \ldots \}$
[F2]	$*$	$\{ \ldots F_i = [\ldots bn(\gamma)(\ldots \gamma_{type\,A/B/C} \ldots) \ldots] , \ldots , F_j = [\ldots bc(\alpha)(\ldots \alpha_{type\,A} \ldots) \ldots] \ldots \}$
[E1a]	$\surd$	$\{ \ldots F_d = [\ldots \gamma_{type\,A/B/C} \ldots] , \ldots , F_e = [\ldots bc(\alpha)(\ldots \alpha_{type\,B} \ldots) \ldots] \ldots \}$
[E1b]	$\surd$	$\{ \ldots F_d = [\ldots \alpha_{type\,B/C} \ldots] , \ldots , F_e = [\ldots bc(\gamma)(\ldots \gamma_{type\,B} \ldots) \ldots] \ldots \}$
[E2]	$*$	$\{ \ldots F_d = [\ldots \gamma_{type\,A/B/C} \ldots] , \ldots , F_e = [\ldots bc(\alpha)(\ldots \alpha_{type\,A} \ldots) \ldots] \ldots \}$
[E3a]	$*$	$\{ \ldots F_d = [\ldots \gamma_{type\,A/B/C} \ldots] , \ldots , F_e = [\ldots \alpha_{type\,C} \ldots] \ldots \}$, if γ c-commands α regardless of the attachment choice
[E3b]	$*$	$\{ \ldots F_d = [\ldots \alpha_{type\,A/B/C} \ldots] , \ldots , F_e = [\ldots \gamma_{type\,C} \ldots] \ldots \}$, if α c-commands γ regardless of the attachment choice
[E4]	$*$	$\{ \ldots F_d = [\ldots \alpha_{type\,A} \ldots] , \ldots , F_e = [\ldots bn(\gamma)(\ldots \gamma_{type\,A/B/C} \ldots) \ldots]$

Figure 1
Rule patterns for binding constraint verification on fragmentary syntax.

[F1]	BP B of α / γ is satisfied	γ does not *locally* bind α and α does not *locally* bind γ
[F2]	BP A of α is violated	γ does not *locally* bind $\alpha \vee \gamma$ does not c-command α
[E1a]	BP B of α is satisfied	γ does not *locally* bind α
[E1b]	BP B of γ is satisfied	α does not *locally* bind γ
[E2]	BP A of α is violated	γ does not *locally* bind α
[E3a]	BP C of α is violated	γ c-commands α
[E3b]	BP C of γ is violated	α c-commands γ
[E4]	BP A of α is violated	γ does not c-command α

Figure 2
Binding-theoretic background of the rule patterns. (BP = binding principle)

become evident, the above basic set of patterns might suffice when the degree of fragmentation of the parsing results is low.

Discussion of some examples will explain these patterns in more detail.

Rule pattern [F1]

$$\surd \quad \{\ldots F_i = [\ldots bc(\gamma)(\ldots \gamma_{type\,B} \ldots) \ldots] , \ldots , F_j = [\ldots bc(\alpha)(\ldots \alpha_{type\,B} \ldots) \ldots] \ldots \}$$

is applicable in cases where two nonreflexive (type B) pronouns α and γ are contained in different surface structure fragments, and, in addition, each fragment contains the binding category (*bc*) of the respective pronoun. Under these conditions, the coindexing of the two pronouns is admissible since, in any possible syntactic reading, it cannot be the case that one of the pronouns *locally* binds the other; that is, the applicable binding principle B will be satisfied in any case. Typical instances are structurally ambiguous adverbial clauses:

(6) The president left after he had spoken because he was tired.

Under the assumptions that the parser lacks the knowledge necessary to structurally disambiguate the *because* clause (which may be interpreted as an adverb of either the main or the *after* clause) and that the syntactic fragments of both adverbial clauses are correctly determined, rule pattern [F1] becomes applicable since, for both pronouns,

the binding category, which is the topmost S node of the respective adverbial clause, is contained in the respective fragment.

Rule pattern [F2]

$$* \quad \{\ldots F_i = [\ldots bn(\gamma)(\ldots \gamma_{type\ A/B/C} \ldots) \ldots] , \ldots , F_j = [\ldots bc(\alpha)(\ldots \alpha_{type\ A} \ldots) \ldots] \ldots\}$$

applies in cases where a reflexive pronoun α occurs in a syntactic fragment that contains its binding category: any candidate γ of arbitrary binding-theoretic type (A, B, or C) that occurs in a different fragment containing its branching node (bn) may be ruled out on configurational grounds since it is impossible to structurally conjoin the two fragments in such a way that γ, as required by binding principle A of α, locally binds α. Here, the condition that the binding category be present in the anaphor's fragment ensures that, in case this fragment is subordinated under the candidate's fragment, no relation of *local* binding holds; on the other hand, the condition that the branching node be present in the candidate's fragment rules out, in the opposite case, the possibility of establishing a c-command relation. [F2] is applicable in the case of example (5). Since the relative clause contains a reflexive pronoun (taken as anaphor α) and the respective binding category, it matches the fragment F_j; similarly, the main clause instantiates F_i with respect to any of its type C occurrences (taken as candidates γ). Hence, according to the prediction of [F2], the immediate, *constructive*[14] coindexing of the reflexive pronoun with any of the candidates occurring in the main clause is ruled out.

For certain adjacent syntactic fragments, the parsing result may comprise additional information about immediate or transitive *embedding*. Based on this evidence, further rule patterns may become applicable (F_d = dominating fragment, F_e = embedded fragment):

Rule patterns [E1a] and [E1b]

$$\surd \quad \{\ldots F_d = [\ldots \gamma_{type\ A/B/C} \ldots] , \ldots , F_e = [\ldots bc(\alpha)(\ldots \alpha_{type\ B} \ldots) \ldots] \ldots\}$$
$$\surd \quad \{\ldots F_d = [\ldots \alpha_{type\ B/C} \ldots] , \ldots , F_e = [\ldots bc(\gamma)(\ldots \gamma_{type\ B} \ldots) \ldots] \ldots\}$$

are the (enhanced) counterparts of pattern [F1]. If the fragment of the type B anaphor α is subordinated, coindexing with an outside candidate γ (here: arbitrarily of type A, B, or C) is admissible. If, on the other hand, the fragment of the type B (or type C) anaphor α is known to be the dominator, a candidate γ of type B that occurs in a fragment containing its binding category is configurationally permitted.[15] Hence, because of the additionally available embedding information, it is possible to relax the demands on the dominating fragment, which is no longer required to contain the binding category of the respective occurrence. Typical cases in which [E1a] and [E1b] apply are instances of structurally ambiguous relative clauses. In example (5), since the (embedded) relative clause fragment contains the binding category of the nonreflexive (type B) pronoun occurrences (taken as anaphors α), fragment F_e of rule [E1a] is instantiated; moreover, trivially, the (dominating) main clause instantiates F_d

14 While binding principle A constructively demands the existence of at least one local binder, it does not preclude further, possibly nonlocal coindexings (cf. the example in Section 4.1). In this sense, the application of [F2] is confined to the *constructive* search for the antecedent required to fulfill binding principle A.

15 In the case of [E1b], the *anaphor* (i.e., the occurrence to be constructively resolved) occurs in the dominating fragment. Since γ cannot be a local binder of α, the occurrence in the dominating fragment is not allowed to be of type A (cf. the remarks on constructive coindexing in footnote 14). Hence, since α and γ are not interchangeable, [E1a] and [E1b] look slightly different.

with respect to any of its (type C) occurrences (taken as candidates γ). Hence, [E1a] applies, licensing the respective coindexings.

Rule pattern [E2]

$$* \quad \{\dots F_d = [\dots \gamma_{type\,A/B/C} \dots] , \dots , F_e = [\dots bc(\alpha)(\dots \alpha_{type\,A} \dots) \dots] \dots \}$$

is the counterpart of pattern [F2]. Under the condition that the anaphor's fragment is known to be subordinated, the restriction that the candidate's fragment contain the respective branching node can be dropped; the presence of the reflexive pronoun's binding category in the embedded fragment proves to be sufficient for ruling out the candidate as the constructive antecedent required according to binding principle A. Again, applied to example (5), [E2] rules out the constructive coindexing of the reflexive pronoun with any candidate occurring in the main clause.

Rule patterns [E3a] and [E3b]

$$* \quad \{ \dots F_d = [\dots \gamma_{type\,A/B/C} \dots] , \dots , F_e = [\dots \alpha_{type\,C} \dots] \dots \},$$
if γ c-commands α regardless of the attachment choice

$$* \quad \{\dots F_d = [\dots \alpha_{type\,A/B/C} \dots] , \dots , F_e = [\dots \gamma_{type\,C} \dots] \dots \},$$
if α c-commands γ regardless of the attachment choice

formally characterize a particular case in which binding principle C is violated: if a type C expression occurs in the embedded fragment, and, in addition, it is known that the other occurrence will c-command the type C expression regardless of the attachment choice, then this coindexing can be definitively ruled out since, in any case, binding principle C will be violated. Typically, these rules apply if the expression occurring in the dominating fragment holds the structurally prominent role of the syntactic subject.

Rule pattern [E4]

$$* \quad \{\dots F_d = [\dots \alpha_{type\,A} \dots] , \dots , F_e = [\dots bn(\gamma)(\dots \gamma_{type\,A/B/C} \dots) \dots]$$

deals with another generic case of binding principle A violation. If the type A pronoun α occurs in the dominating fragment, and, in addition, the subordinated fragment contains the branching node of a constructive candidate γ of arbitrary binding-theoretic type, this candidate can be ruled out since, in any possible case of structural recombination, γ will not c-command α; in particular, this implies that, as required by binding principle A, γ does not (locally) bind α. Since the requirement on the constructive candidate's fragment is weak, pattern [E4] applies in virtually any case in which a reflexive pronoun occurs in a dominating fragment.

In general, there may be more than one rule pattern applying to a certain configuration.[16] However, the set of patterns is *consistent* in the sense that, whenever this situation arises, the predictions of all applicable rules are identical.

There are two further rule patterns that match certain syntactic configurations in which a coindexing would violate the i-within-i condition (cf. Figure 3). Both patterns are abstractions over cases of fragment embedding in which the root of the dominating fragment constitutes one of the relevant occurrences. Thus, the scope of the i-within-i patterns is rather restricted. As an example, if there is a dominating NP fragment (constituting an antecedent candidate γ) and a locally ambiguous PP fragment containing an anaphor α, [IEa] rules out coindexing the anaphor's NP with the overall NP.

16 For example, [F2] as well as [E2] in the case of the reflexive pronoun in example (5).

[IEa] $*\quad \{ \ldots F_d = [\gamma(\ldots)] , \ldots , F_e = [\ldots \alpha(\ldots) \ldots] \ldots \}$
[IEb] $*\quad \{ \ldots F_d = [\alpha(\ldots)] , \ldots , F_e = [\ldots \gamma(\ldots) \ldots] \ldots \}$

Figure 3
Rule patterns for i-within-i condition verification on fragmentary syntax.

4. Anaphor Resolution on Fragmentary Syntax: The ROSANA System

Based on the above set of rule patterns, an anaphor resolution algorithm can be designed that achieves robustness against fragmentary syntactic descriptions according to the deficient description model.

4.1 The ROSANA Algorithm

Figure 4 describes the ROSANA algorithm. By applying a set of *restrictions* (Step 1) prior to a set of *preferences* (Step 2), this algorithm follows Carbonell and Brown's (1988) fundamental strategy by means of which the candidate set is narrowed down as early as possible. In Step 3, the actual *selection* of antecedents takes place. Among the strategies to be applied are restrictions (e.g., morphosyntactic and lexical congruence, disjoint reference conditions) and a plethora of preference factors (subject/topicalization salience, syntactic obliqueness, recency, cataphor penalty, parallelism [inertia of syntactic function]). Since the goal is to design an anaphor resolution *algorithm*, the choice is restricted to strategies that are operational.

With respect to syntactic disjoint reference, the central goal of robustness against fragmentary syntax is achieved in Steps 1(b) and 3(b). As described above, if the considered occurrences are situated in different syntactic fragments, the rule patterns come into play; the actual set of patterns to be applied depends on whether or not it is known that one of the fragments is embedded in the other. Patterns labeled "$*$" are used to eliminate candidates (Steps 1(b)iv and 1(b)v). Patterns marked "$\surd$" are used to *definitively* admit candidates (Step 1(b)vi), contrasting with *heuristic* admittance (Step 1(b)vii), which entails a decrement of the plausibility score in Step 2(a).

One subtlety taken into account is **interdependency** between different antecedent decisions (cf. Step 3). In particular, decision interdependency may arise because of the transitivity of the coindexing relation. As illustrated by the following example, even regarding *inter*sentential anaphora, antecedent decisions that *individually* comply with the disjoint reference conditions may *collectively* induce a violation:

(7) $*$Gropius$_i$ discusses the plans with Behrens$_j$. He$_i$ meets him$_i$ in Dessau.

For each of the two pronouns, candidate *Gropius* is configurationally admissible. In a formal sense, however, the binding principles state restrictions on (intrasentential) index *distributions* rather than on single anaphor-candidate pairs in isolation: in the example, binding principle B of the pronoun *him* is *transitively* violated. In explicitly checking for the binding-theoretic admissibility of transitively induced coindexings, the algorithm guards against such cases (Step 3(b)). However, care must be taken not to apply the binding restriction for reflexives constructively in this test since, as illustrated in the following example, besides demanding constructively one local binder, binding principle A does not rule out *further* nonlocal coindexings:

(8) Gropius$_i$ admits that he$_i$ shaves himself$_i$.

1. *Candidate filtering*: For each anaphoric NP α, determine the set of admissible antecedents γ:
   (a) verify morphosyntactic or lexical agreement with γ;
   (b) if the antecedent candidate γ is intrasentential:
   - if α and γ belong to the same syntactic fragment, then verify that
     i. the binding restriction of α is constructively satisfied,
     ii. the binding restriction of γ is not violated,
     iii. no i-within-i configuration results;
   - else (α and γ belong to different syntactic fragments) *try the rule patterns*:
     iv. if one of the patterns [E2], [E3a], [E3b], [E4], or [F2] is matched, then some binding restrictions are violated,
     v. else if one of the two i-within-i rule patterns applies, then some binding restrictions are violated,
     vi. else if pattern [E1a], [E1b], or [F1] applies, then the binding restrictions of α and γ are satisfied,
     vii. else (*no rule pattern applies*) assume heuristically that the binding restrictions of α and γ are satisfied;

   (c) if α is a type B pronoun, antecedent candidate γ is intrasentential, and, with respect to surface order, γ *follows* α, verify that γ is *definite*.
2. *Candidate scoring and sorting*:
   (a) For each remaining anaphor-candidate pair (α_i, γ_j): based on a set of preference heuristics, determine the numerical plausibility score $v(\alpha_i, \gamma_j)$. If the binding-theoretic admissibility was approved *heuristically* in step 1(b)vii, then reduce the plausibility score $v(\alpha_i, \gamma_j)$ by a constant value;
   (b) for each anaphor α: sort candidates γ_j according to decreasing plausibility $v(\alpha, \gamma_j)$;
   (c) sort the anaphors α according to decreasing plausibility of their respective best antecedent candidates.
3. *Antecedent selection*: Consider anaphors α in the order determined in Step 2(c). Suggest antecedent candidates $\gamma_j(\alpha)$ in the order determined in Step 2(b). Select $\gamma_j(\alpha)$ as candidate if there is no interdependency, that is, if
   (a) the morphosyntactic features of α and $\gamma_j(\alpha)$ are still compatible,
   (b) for all occurrences $\delta_{\gamma_j(\alpha)}$ and δ_α the coindexing of which with $\gamma_j(\alpha)$ and (respectively) α has been determined in the *current* invocation of the algorithm: the coindexing of $\delta_{\gamma_j(\alpha)}$ and δ_α, which results transitively when choosing $\gamma_j(\alpha)$ as antecedent for α, violates neither the binding principles nor the i-within-i condition; that is,
   - if $\delta_{\gamma_j(\alpha)}$ and δ_α belong to the same syntactic fragment, then, for both occurrences, verify the respective binding conditions and the i-within-i condition according to steps 1(b)ii and 1(b)iii,
   - else if $\delta_{\gamma_j(\alpha)}$ and δ_α belong to different syntactic fragments, then proceed according to steps 1(b)iv, 1(b)v, 1(b)vi, and 1(b)vii (with the exception of the rule patterns [F2], [E2], and [E4], by means of which binding principle A is *constructively* verified).

   (The case $\delta_{\gamma_j(\alpha)} = \gamma_j(\alpha) \wedge \delta_\alpha = \alpha$ does not need to be reconsidered.)

Figure 4
The ROSANA anaphor resolution algorithm.

The same distinction is drawn in Step 1(b): whereas, regarding the anaphor, the binding restriction is verified in the strong, constructive sense (Step 1(b)i), the candidate's restriction is applied in its weak version (Step 1(b)ii). In the rule patterns for the fragmentary case, this subtlety is reflected implicitly in the sense that only regarding occurrence α (taken as the anaphor to be constructively resolved) is the strong version of binding principle A checked; hence, in the interdependency test Step 3(b), patterns [F2], [E2], and [E4] are not taken into consideration.

4.2 Implementation: The ROSANA System

Based on the algorithm described in Figure 4, the ROSANA anaphor resolution system has been implemented (Stuckardt 2000). In primarily aiming at determining the coreference classes of nonzero linguistic expressions that specify entities,[17] the scope of ROSANA corresponds to the coreference task of the Message Understanding Conferences (cf. Hirschman 1998). ROSANA handles a broad range of entity-specifying expressions—in particular, ordinary, possessive, reflexive/reciprocal, and relative pronouns, definite NPs, and names. The ROSANA system has been implemented in Common Lisp. In an evaluation on a set of news agency press releases (cf. Section 6), the runtime of the ROSANA system (without parser) was 165 tokens per second on a Pentium PC.

The FDG parser for English developed by Järvinen and Tapanainen (1997) has been chosen as the syntactic preprocessor.[18] In giving robustness and processing speed priority over normativity and syntactic coverage of the underlying grammar, the parser meets the requirements on a preprocessor for robust anaphor resolution on unrestricted texts.[19] Regardless of the typical parsing problems like structurally ambiguous or grammatically incorrect input, the parser always yields a result, comprising one or more syntactic fragments that cover the analyzed sentence. Hence, the parser is regarded to be an ideal associate of robust anaphor resolution approaches that follow the deficient description model.

4.3 Anaphoric Occurrences and Antecedent Candidates

In the ROSANA system, the above algorithm is supplemented by a set of strategies for *identifying* occurrences (linguistic expressions that specify entities) and *classifying* them as anaphors to be resolved and/or as possible antecedent candidates. The criterion for the identification of specifying expressions is based on part of speech (as determined by the FDG parser) and syntactic context (Stuckardt 2000, page 249).[20] Also, the decision of anaphoricity is based on evidence regarding the syntactic context: generally, occurrences of all three binding-theoretic types (A, B, C) are taken to be anaphoric;[21] however, there are some classes of anaphoric occurrences that may, in certain cases, be interpreted in advance (outside the ROSANA core algorithm) by purely structural means (e.g., relative pronouns or occurrences induced by heads of appositions). For narrowing down the search space, occurrences of type A (reflexive and reciprocal pro-

17 In contrast to expressions that, for example, specify events.

18 Since the parser generates dependency descriptions rather than constituent structure (to which the formal definitions of the above GB-theoretic statement of syntactic disjoint reference refer), ROSANA applies a preprocessor that reconstructs the structural (e.g., subject-object) asymmetries of constituency that are vital to the verification of the disjoint reference conditions.

19 According to Järvinen and Tapanainen (1997), the FDG parser processes an average of 350 words per second on modest hardware (Pentium PC, 166 MHz).

20 Syntactic context plays a role, for example, in deciding whether the expression *her* is a possessive or nonpossessive pronoun.

21 In the case of nonpronominal NPs, it proves to be difficult to decide algorithmically (e.g., based on information about the determiner) whether or not a new discourse referent is introduced.

nouns) are not taken into account as antecedent candidates, since the existence of a cospecifying alternative in the same clause remains guaranteed.

4.4 Congruence Conditions

In Step 1(a) of the ROSANA algorithm, the details regarding the congruence restrictions are left unspecified. In the ROSANA system, depending on the specific type of anaphoric expression, different morphosyntactic or lexical agreement conditions are employed. For names, for example, a partial matching of the antecedent and anaphor expressions (in the sense of surname identity) is considered sufficient. Regarding third person pronouns (including possessives and reflexives/reciprocals), congruence of the morphological features *number* and *person* is considered mandatory; however, congruence of the *gender* attribute is taken to be optional, since, on one hand, there are some well-known exceptions, and, on the other hand, the available grammatical gender information is not always correct.[22] In any case, candidates that also match the gender attribute of the anaphor are preferred.

4.5 Salience Factors and Weights

As in the approaches of Lappin and Leass (1994) and Kennedy and Boguraev (1996), **weighted salience factors** are employed for scoring and choosing among the candidates that remain after restriction application (cf. Step 2(a) of the ROSANA algorithm). Since the goal is to develop an *operational* approach for unrestricted texts, the choice is restricted to factors relying on information available in the scenario of knowledge-poor processing, that is, without extensive semantic domain modeling.

In the ROSANA system, the following factors are employed: SYR (contributed by occurrences with identical syntactic function), EEP (occurrences realized in the syntactic position of existential emphasis), SUP (syntactic subject), PGP (possessive pronouns, saxonian genitives, and genitive attributes), DOP/IOP/APP (salience of direct/indirect objects and adverbial PPs, respectively), KAM (negative preference of cataphoric resumptions), SDM (sentence recency; i.e., a factor of negative salience to be multiplied with the sentence distance between anaphor and antecedent), WDM (word recency). The main part of Table 1 indicates which subset of factors is used for scoring the candidate set of which class of anaphoric expression (DNOM = definite NP, PER{1,2,3} = first/second/third person pronouns, POS{1,2,3} = first/second/third person possessives, RELA = relative pronouns, REFL = reflexive/reciprocal pronouns).

The assignment of the factors and the choice of the weights (shown in the lower part of the table) have been determined by a series of refinement experiments on a training corpus of 31 news agency press releases (11,808 words, 471 pronouns) for which key data were provided manually. As a proper base for the goal-directed refinement of the factor assignments and weights, the interpretation results were scored according to two of the formal evaluation disciplines that will be defined in Section 5, namely, determination of coreference classes (model-theoretic scoring) and nonpronominal anchors.[23] The factor/weight relations determined by Lappin and Leass (1994) for third person pronouns were taken as the initial clue. For the other types of anaphoric expressions, sets of weighted factors were assigned based on an analysis of the referential context of typical occurrences in the training corpus. For example, the

22 This is partly due to lexical ambiguity (homonymy), or, regarding names, due to lack of the respective lexical information.

23 Since, for these disciplines, two-dimensional (precision/recall) measures are defined, there may, in general, be multiple (pareto-)optimal factor/weight assignments. Instead, one may refer to a combined (weighted) scoring scheme, such as the F measure employed in the MUC evaluations.

Table 1
Salience factors and weights for different types of anaphoric expressions.

Anaphor type	SYR	EEP	SUP	PGP	DOP	IOP	APP	KAM	SDM	WDM
DNOM								+	+	
NAME								+		
PER3	+	+	+	+	+	+	+	+	+	
PER2	+	+	+	+	+	+	+	+	+	
RELA										+
POS3	+		+					+	+	
POS2	+		+					+	+	
REFL										
PER1			+					+		+
POS1			+					+		+
Weights	20	15	15	13	10	5	5	125	25	25

corpus study revealed that first person nonpossessive and possessive pronouns typically resume discourse referents instantiated by nearby antecedents that occur in the syntactic subject role; consequently, the factors SUP, WDM, and KAM were chosen.[24] During a series of variation and evaluation runs in which some of the factors were systematically deactivated, these initial assignments were empirically validated.[25]

Finally, a series of experiments with the factor weights were carried out. Clearly, the absolute size of the weights is irrelevant. However, some conditions that seem to govern the *relative* size of the factor weights were determined/confirmed during the training runs. For example, in locally varying the assigned weights in such a way that individual $\geq$ relations of the syntactic function hierarchy SUP $\geq$ DOP $\geq$ IOP $\geq$ APP were violated (e.g., by setting SUP $=$ 10 $<$ DOP $=$ 15), it was experimentally verified that the original weight relations yield better results. Further findings are: SDM $>$ SYR (syntactic parallelism induces *local* preferences only); SYR $>$ SUP (if an anaphor occurs in a syntactic role other than subject, then candidates with that same role are preferred to candidates in subject role); KAM large (cataphoric resumptions are heavily penalized). For the most part, these results coincide with, or provide further support for, similar findings by Lappin and Leass (1994, page 549).[26]

As table 1 makes evident, the salience factors proper (determined by syntactic role) are employed only in the case of pronominal anaphors. Most importantly, there is a striking difference between PER3 and POS3: whereas, in the former case, a hierarchy of syntactic roles (i.e., salience factors with decreasing weights) is referred to, in the latter case, only the factors of subject preference and syntactic parallelism are employed because it turned out that possessive pronouns tend to cospecify with antecedents that are either syntactic subjects or, again, possessive pronouns. Relative pronouns are considered to be an exception: since, in most cases, they take their antecedents in the nearest vicinity, the word recency factor proved to be sufficient.

24 Through this preference, cases are accounted for, too, in which appropriate third person antecedents occur outside a passage of quoted speech containing the first person pronoun. In particular, this renders possible the determination of nonpronominal anchors (cf. Section 5.3) for first person pronouns.

25 Whereas, in large part, the experiments confirmed the factor assignments for third person pronouns suggested by Lappin and Leass, it turned out that the training corpus did not contain a sufficient number of occurrences realized in the syntactic position of existential emphasis for evaluating the contribution of the EEP factor. This should be addressed by further experiments on larger corpora.

26 One divergence regards the size of the syntactic parallelism factor SYR. According to the experimental results, SYR should be larger than SUP. However, Lappin and Leass determined that SYR should be just large enough to offset the preference for subjects over accusative objects, that is, SYR + DOP $>$ SUP. This issue should be addressed by further experiments.

5. Enhancing the Evaluation Methodology for Anaphor Resolution Systems

For a proper evaluation of ROSANA, appropriate evaluation measures have to be chosen. In the following discussion, it will be advocated that, to obtain results that are expressive from the developer's as well as the application's point of view, several evaluation disciplines should be considered.[27]

5.1 Model-Theoretic Coreference Scoring

Vilain et al. (1996) developed the **model-theoretic scoring scheme** according to which precision and recall values are computed by a formal alignment of the coreference equivalence classes of system output and intellectually gathered key data. Basically, *precision errors* correspond to nontrivial partitions of system-generated equivalence classes induced by key equivalence classes, whereas *recall errors* correspond to nontrivial partitions of key equivalence classes induced by system-generated equivalence classes.

Formally, let R^s and R^k be the coreference relations computed by the anaphor resolution system and specified by the key, respectively; moreover, let $[R^s]$ and $[R^k]$ be the respective sets of equivalence classes. Furthermore, with $C^s \in [R^s]$ and $C^k \in [R^k]$, let $C^s \cap O^k$ and $C^k \cap O^s$ be the equivalence classes (sets of occurrences) obtained by *restricting* the original equivalence classes to the sets of occurrences O^k and O^s over which the relations R^k and R^s, respectively, are defined;[28] analogously, let $\Phi(C^s, R^k)$ and $\Phi(C^k, R^s)$ be the equivalence relations that result by *restricting* the original relations R^k and R^s to the occurrences contained in the equivalence classes $C^s \in [R^s]$ and $C^k \in [R^k]$, respectively (cf. the discussion in Section 5.2.1).[29] In addition, let $[\Phi(C^k, R^s)]$ and $[\Phi(C^s, R^k)]$ be the sets of equivalence classes of the restricted relations. The precision and recall measures are computed by summing over the sets of equivalence classes of system response and key classes, respectively. For each class C, there is a maximum of $|C \cap O| - 1$ correct contributions; the actual number of errors, which equals the number of equivalence classes $|[\Phi(C, R)]|$ of the restricted relation minus 1, has to be deducted. Hence, one obtains the measures

$$P_{co} := \frac{\sum_{C^s \in [R^s]} (|C^s \cap O^k| - |[\Phi(C^s, R^k)]|)}{\sum_{C^s \in [R^s]} (|C^s \cap O^k| - 1)}$$

$$R_{co} := \frac{\sum_{C^k \in [R^k]} (|C^k \cap O^s| - |[\Phi(C^k, R^s)]|)}{\sum_{C^k \in [R^k]} (|C^k \cap O^s| - 1)}$$

5.2 Scoring from the Developer's Point of View

The above precision and recall measures refer to a formal, mathematical property of the structure of the results computed by coreference resolution systems, namely, the disjoint partitioning of the set of occurrences in equivalence classes of cospecifying

27 See also Mitkov 2001, in which, independently, similar proposals regarding the separate evaluation of anaphor resolution system components have been made. There are some further important contributions of this paper that can be regarded as complementary to the work presented below, particularly the definition of formal evaluation measures for determining the *decision power* and the *relative importance* of individual salience factors.

28 This means that $R^k \subseteq O^k \times O^k$ and $R^s \subseteq O^s \times O^s$.

29 To put it formally: $\Phi(C^s, R^k) = R^k \cap (C^s \times C^s)$ and $\Phi(C^k, R^s) = R^s \cap (C^k \times C^k)$.

entities. Whereas the definitions are appealing from the point of view of theoretical elegance, from the system developer's perspective they display certain shortcomings.

5.2.1 Supporting the Optimization of Component Algorithms. A first point of criticism is the lack of expressiveness regarding the different typical subproblems to be solved by coreference resolution systems. One particular subtask that is usually handled by a preprocessor is the identification of relevant occurrences, that is, entity-specifying linguistic expressions. As a suitable base for evaluating and optimizing this module, it seems to be adequate to dedicate separate evaluation measures, the definition of which, in this case, is straightforward: let O^s and O^k be the sets of occurrences computed/specified by the coreference resolution system/key, then set

$$P_{oc} := \frac{|O^s \cap O^k|}{|O^s|}$$

$$R_{oc} := \frac{|O^s \cap O^k|}{|O^k|}$$

To ensure the expressiveness of the totality of evaluation measures, it is essential that they be *decoupled* from each other in the sense that errors at one stage of processing are *exclusively* reflected in the respective evaluation measure. Regarding the possible effects of precision and recall errors of occurrence identification on model-theoretic coreference scoring, this requirement is met by referring, in the definitions given in Section 5.1, to the *restricted* classes $C^s \cap O^k$ and $C^k \cap O^s$, and to the *restricted* relations $\Phi(C^s, R^k)$ and $\Phi(C^k, R^s)$. Without this refinement of the model-theoretic measures, there may be cases of scoring anomalies. If the cardinalities in the above definitions were determined by referring to the original (unreduced) equivalence classes and relations, each additional occurrence $o^s \in C^s \backslash O^k$ or $o^k \in C^k \backslash O^s$ would lead to an (incorrect) increase in precision or recall, respectively, because, trivially, these sets of occurrences are not partitioned by the relations R^k or R^s, respectively.

5.2.2 Supporting the Refinement of Preferential Factors. As shown in Table 1, the set of relevant salience factors depends on the specific type of anaphoric expression. Hence, the evaluation measures defined so far are considered insufficient for an optimization of factor assignments and weights. From the system developer's perspective, there is a need for fine-grained information that distinguishes between different classes of anaphoric expressions.

As will become evident during evaluation of the ROSANA system in Section 6, another reason for differentiating between types of anaphoric expressions is the lack of expressiveness of model-theoretic scoring regarding the interpretation quality achieved for *pronouns* (i.e., for the class of anaphors that is, from the perspective of typical applications, of central importance).

5.3 Scoring from an Application's Point of View: Nonpronominal Anchors

Regarding the requirements of typical applications,[30] the task of pronoun interpretation may be defined as determining a suitable nonpronominal substitute, that is, a *nonpronominal* antecedent rather than an *arbitrary* cospecifying antecedent that again might be a pronoun.

30 For example, the MUC information extraction task proper (Scenario Template), or the classical quantitative, dictionary-based content analysis of the social sciences.

According to model-theoretic coreference scoring, no distinction is made between pronominal and nonpronominal antecedents: what is relevant are the sizes of the matching fractions of equivalence classes rather than the presence of correct **nonpronominal anchors** suitable as substitute expressions. There are at least two reasons why the task of identifying correct nonpronominal substitutes is considerably harder than the task of finding an arbitrary correct antecedent. First, there is focus-theoretic evidence: typically, entities specified by pronouns are in focus[31] and, hence, most probably the antecedents of subsequent pronouns, which tend to resume the currently focused entity.[32] Second, a technical argument applies: the cospecification relation between a pronominal anaphor and a nonpronominal representative is algorithmically determined by a nonempty *chain* of antecedent decisions that may, in general, be long:

$$\text{Gropius} \xleftarrow{-} \text{he} \xleftarrow{+} \text{he} \xleftarrow{+} \text{him} \xleftarrow{+} \text{him}$$

Whereas the hypothetical single error (indicated by "$-$") implies incorrect nonpronominal anchors for all pronouns, according to the model-theoretic measure the precision amounts to 0.75. Consequently, the evaluation scheme defined so far should be supplemented by a measure that is expressive with respect to the application-relevant task of identifying nonpronominal anchors.

To derive a suitable formal evaluation measure, one may start with the observation that the relevant linguistic entities to be resolved are pronominal occurrences P for which nonpronominal anchor occurrences A have to be identified. Basically, an anchor A shall be considered a *correct* substitute for P if and only if A and P belong to the same equivalence class of the key coreference relation. From a theoretical point of view, this definition must be considered simplistic given the well-known examples of opaque (intensional) contexts in which the substitution of coreferring expressions is not a truth-preserving operation. These cases, however, are rare and do not seem to play a role in typical application scenarios of pronoun interpretation algorithms.[33] Hence, the simple definition will be employed, which, in addition, entails the advantage that the key data provided for model-theoretic coreference scoring suffices for the scoring of nonpronominal substitutes.

Let (P, A) be a pair consisting of a pronominal occurrence P and the anchor occurrence A determined by the anaphor resolution system. (If, for P, no substitute has been determined, then A is considered empty.) A suitable base for scoring is obtained by classifying the pairs (P, A) according to the scheme described in Table 2, by which a total of seven pairwise disjoint sets is defined. The classification depends on (1) whether P and/or A are tagged, in the key, as valid (entity-specifying) occurrences, (2) whether A is nonempty, and (3) whether, in case A is nonempty and both A and L are valid occurrences, A and L cospecify in the key. According to the above definition, only the pairs fulfilling condition (3) (which, hence, constitute the set o_{++}) are considered to be correct solutions.

31 Care should be taken not to confound two different notions of focus here. In terms of the classical topic-focus distinction, one would say that pronouns tend to specifiy entities constituting the *topic*.

32 Compare, for example, the predictions of the centering theory (Grosz, Joshi, and Weinstein 1995).

33 There is another argument that supports the choice of the simple definition. Probably the best algorithmic strategy for determining correct anchors is the selection of the first nonpronominal cospecifying occurrence that topologically precedes the pronoun to be resolved (cf. the above decision chain argument). If, however, the distance between the determined substitute and the pronoun is small, from the point of view of conversational pragmatics it is implausible that the intension of the substitute occurrence does not match the (possibly opaque) context, since, otherwise, human readers are expected to be misled as well (Stuckardt 2000, page 240).

Table 2
Classification and scoring of nonpronominal anchors.

Set	Scoring	Definition
o_{++}	correct	P and A belong to the same key equivalence class
o_{+-}	incorrect	P and A belong to different key equivalence classes
$o_{+?}$	incorrect	P, but not A, corresponds to a key occurrence
$o_{+\_}$	empty	P corresponds to a key occurrence, no anchor A determined
o_{+*}	empty	P corresponds to a key occurrence, no anchor A determined, cospecification of P is marked as optional in key
$o_{?+}$	incorrect	P does not correspond to a key occurrence
$o_{?\_}$	empty	P does not correspond to a key occurrence, no anchor A determined

Again, the requirement of mutually decoupling the evaluation measures should be fulfilled. Regarding the errors made during the identification of specifying occurrences, this goal may be met by basing the measures on the sets o_{++}, o_{+-}, $o_{+?}$, $o_{+\_}$, and o_{+*} that constitute the cases in which the base entity to be decided upon, namely, the *pronoun* occurrence P, has been determined in compliance with the key. By further drawing the usual distinction between precision and recall, according to which, in the latter case, one must take into account empty anchors A as well, one obtains the following definitions:[34]

$$P_{na} := \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}|}$$

$$R_{na} := \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}| + |o_{+\_}|}$$

Since errors in the occurrence identification are excluded from measurement at this stage of evaluation, and, moreover, it is assumed that there are no errors regarding the classification of occurrences as decision-relevant entities (i.e., *pronouns*),[35] it follows that, in any case, $P \geq R$. However, the characteristic trade-off relation between precision and recall holds anyway. If the assignment of nonpronominal anchors is confined to highly plausible decisions, whereas the set $o_{+\_}$ will be larger, the sets o_{+-}, $o_{+?}$, and, expectedly to a lesser extent, o_{++} will be smaller, thus typically yielding higher precision and lower recall. Vice versa, if more decisions are performed, $o_{+\_}$ will decrease in size, but o_{+-}, $o_{+?}$, and, expectedly to a lesser extent, o_{++} will be larger, thus tending to higher recall and lower precision. The special case $P = R$ holds if the set $o_{+\_}$ is empty, that is, if there are no open decisions.[36]

34 In generalizing the handling of optional coreferences (as originally specified in the coreference task definition [Hirschman 1998] with respect to model-theoretic scoring), unresolved pronouns whose antecedent link is marked as optional in the key (i.e., the elements of the set o_{+*}) are not taken into account in the recall measure of the nonpronominal anchor discipline.

35 The latter simplification is unproblematic because, under the condition that an expression is a valid occurrence, the decision whether it represents a *pronoun* is trivial.

36 The one-dimensional accuracy measure that is typically employed in the evaluation of pronoun resolution algorithms (for example, Lappin and Leass [1994] or Kennedy and Boguraev [1996]) implicitly relies on the fact that *all* pronouns are resolved. Under this condition, distinguishing between precision and recall becomes unnecessary. Hence, the definition of the two-dimensional measure (P_{na}, R_{na}) must be considered a *generalization* of the conventional accuracy measure. Employing the refined precision/recall distinction even in the case of the arbitrary antecedent discipline may be appropriate when evaluating anaphor resolution systems that aim at achieving high precision by leaving some of the decisions open.

6. Evaluation of the ROSANA System

Figure 5 shows the evaluation results of the ROSANA system on a corpus of 35 news agency press releases, comprising 12,904 words and 479 pronouns.[37] The evaluation, which was performed according to the enhanced set of evaluation measures, took place under application conditions, that is, without an a priori manual correction of orthographic or syntactic errors.

6.1 Entity-Specifying Occurrences

The upper part of Figure 5 displays the score on the discipline of identifying entity-specifying occurrences (cf. Section 5.2.1): this subproblem is solved by ROSANA with a (P_{oc}, R_{oc}) performance of (0.94,0.96). Regarding the precision errors, a closer look reveals that approximately 50% of the 243 overgenerated occurrences can immediately be traced back to errors during morphological and syntactic analysis (in particular, there were a number of cases in which adjectives were wrongly classified as nouns, or in which the parsing of a compound NP failed); another 40% are failures of the ROSANA occurrence identification algorithm proper. Regarding recall, fewer than 20% of the missing 150 occurrences are due to errors of the ROSANA occurrence identification algorithm.

With respect to the identification of *pronouns*, the performance of (0.94,0.996) is considerably higher. In this important case, precision errors that were caused by miscategorization of nonreferential occurrences of the expressions *it* and *that* are the main problem. An improvement of approximately 50% may be gained by refining the syntactic analysis, which, at present, fails in certain cases (e.g.) to recognize nonreferential occurrences of *it* as formal subjects.

6.2 Coreference Classes and Immediate Antecedents

According to the model-theoretic coreference scoring scheme defined in Section 5.1, the (P_{co}, R_{co}) performance of ROSANA is (0.81,0.68) (Figure 5, coreference classes). A closer analysis of the correctness of the immediate antecedents makes evident that the actual interpretation quality varies heavily with respect to the type of anaphoric expression.[38] Whereas the precision regarding the antecedent choices for names amounts to 0.94, the performance for the important classes of third person pronouns and possessives is considerably lower (0.71 and 0.76, respectively).[39] Furthermore, as expected, the precision for reflexive/reciprocal pronouns is optimal (1.0) since binding principle A yields tight syntactic bounds that delimit the space of possible antecedents. Regarding definite NPs, the interpretation quality is considerably lower (0.7) because, at present, ROSANA relies on a simple test for lexical recurrence and number agreement and does not employ enhanced techniques for the interpretation of nonpronominal anaphora.[40] Finally, the precision for first and second person pronouns is quite

37 A scoring module has been implemented by which the above-defined evaluation measures are computed. Reference data have been provided by an intellectual annotation of the press release corpus according to the MUC-7 coreference task definition (Hirschman 1998).

38 Regarding the anaphor type abbreviations employed in Figure 5, see Section 4.5.

39 The figures regarding the correctness of immediate *arbitrary* antecedents have been determined according to the precision measure that was originally developed for scoring *nonpronominal* anchors. Since, at least for the most common types of third person pronouns (PER3, POS3), immediate antecedents are determined in virtually any case, the recall figures are almost identical (cf. the discussion in Section 5.3). The precision measure coincides with the accuracy measure employed by, for example, Lappin and Leass (1994) or Kennedy and Boguraev (1996).

40 The results in Figure 5 also indicate that a huge fraction of definite NP occurrences (1,973 + 43) are not assigned an antecedent. This figure, which at first sight seems to be too high, is of the right order of

```
OCCURRENCES:
 - SYS: 243
 - KEY: 150
 - SYS AND KEY: 3831
=> PRECISION: 0.9404
=> RECALL: 0.9623


COREFERENCE CLASSES:

 SYSTEM CLASSES
 - CUTS: 256
 - POSSIBLE: 1334
=> PRECISION: 0.8081

 KEY CLASSES
 - CUTS: 496
 - POSSIBLE: 1572
=> RECALL: 0.6845


IMMEDIATE ANTECEDENTS:

             | PRECIS |  ++  |  +-  |  +?  |  +_  |  +*  |  ?+  |  ?_  |
 ------+------+--------+------+------+------+------+------+------+------+
 PRON | PER3 | 0.7143 |  145 |   48 |   10 |    1 |    0 |   18 |    0 |
      | PE12 | 0.9474 |   18 |    1 |    0 |    7 |    6 |    0 |    0 |
      | POS3 | 0.7634 |  100 |   28 |    3 |    0 |    0 |    0 |    0 |
      | PO12 | 1.0000 |    3 |    0 |    0 |    1 |    1 |    0 |    0 |
      | REFL | 1.0000 |    3 |    0 |    0 |    1 |    0 |    0 |    0 |
      | RELA | 0.7789 |   74 |   18 |    3 |    6 |    0 |    7 |    4 |
      +------+--------+------+------+------+------+------+------+------+
             | 0.7555 |  343 |   95 |   16 |   16 |    7 |   25 |    4 |
 ------+------+--------+------+------+------+------+------+------+------+
 NOMN | DNOM | 0.7014 |  357 |  136 |   16 | 1973 |   43 |   31 |  133 |
      | NAME | 0.9390 |  308 |   15 |    5 |  368 |    5 |    5 |   28 |
      +------+--------+------+------+------+------+------+------+------+
             | 0.7945 |  665 |  151 |   21 | 2341 |   48 |   36 |  161 |
 ------+------+--------+------+------+------+------+------+------+------+

 AVERAGE
=> PRECISION: 0.7808


NONPRONOMINAL ANCHORS:

      | PRECIS | RECALL |  ++  |  +-  |  +?  |  +_  |  +*  |  ?+  |  ?_  |
 -----+--------+--------+------+------+------+------+------+------+------+
 PER3 | 0.6766 | 0.6667 |  136 |   54 |   11 |    3 |    0 |   18 |    0 |
 PE12 | 0.9091 | 0.3846 |   10 |    1 |    0 |   15 |    6 |    0 |    0 |
 POS3 | 0.6641 | 0.6641 |   87 |   39 |    5 |    0 |    0 |    0 |    0 |
 PO12 | 1.0000 | 0.5000 |    2 |    0 |    0 |    2 |    1 |    0 |    0 |
 REFL | 1.0000 | 0.7500 |    3 |    0 |    0 |    1 |    0 |    0 |    0 |
 RELA | 0.7667 | 0.6832 |   69 |   18 |    3 |   11 |    0 |    7 |    4 |
 -----+--------+--------+------+------+------+------+------+------+------+

 AVERAGE
=> PRECISION: 0.7009
=> RECALL: 0.6532
```

Figure 5
Results of ROSANA on the news agency press releases evaluation corpus.

high, too (0.95 and 1.0 for nonpossessives and possessives, respectively), mainly due to the person congruence condition. Regarding pronouns, a closer analysis shows that approximately 30% of precision errors are due to the assignment of incorrect gender attributes during morphological analysis and occurrence identification, and, hence, may be eliminated if additional lexical information becomes available. Another 30% of errors are induced by cases that are beyond the horizon of the heuristic salience-based antecedent ranking in the sense that a theoretically adequate solution would rely on background knowledge that usually is unavailable in unrestricted application contexts.

On one hand, the discussion reveals that refinements should focus on third person pronouns and definite NPs. On the other hand, the results also indicate that, on the basis of the information usually available in knowledge-poor environments, there is little room for further improvement.[41]

6.3 Nonpronominal Anchors

Regarding the task of identifying nonpronominal anchors, the results fall considerably below the figures determined above for the immediate (arbitrary) antecedent case. According to the (P_{na}, R_{na}) measures defined in Section 5.3, the average precision is reduced to 0.70 (compared with 0.76); regarding third person pronouns and possessives, the precision decreases to 0.68 (0.71) and 0.66 (0.76), respectively.

The striking difference with the results of model-theoretic and immediate (arbitrary) antecedent scoring confirms the arguments put forward in Section 5.3 according to which the determination of nonpronominal anchors is considerably harder: whereas, for 306 pronouns, correct immediate (arbitrary) antecedents as well as nonpronominal anchors were determined, there are another 21 cases in which only the former choice proved to be correct (cf. the chain argument). Furthermore, the focus-theoretic argument is supported: out of the selected antecedents of type pronoun, 85.6% were correct, whereas, out of the selected antecedents of type definite NP/name, only 71.3% were correct. Hence, as a proper base for obtaining results that are expressive with respect to the pronoun substitution task, the model-theoretic coreference scoring scheme should be supplemented with the described additional measures.

6.4 Toward a Genre-Specific Assignment of Preference Factors

Since the assignment of the salience factors and their weights has been heuristically optimized on a training corpus of news agency press releases (cf. Section 4.5), the question arises whether, on one hand, these settings are still optimal on the evaluation set of press releases, and, on the other hand, they are optimal on other corpora drawn from different genres and domains as well. Five experiments were conducted to address these topics. The second evaluation corpus consisted of three texts describing the plots of Mozart operas. These texts, which comprise 2,522 words and 236 pronouns, were considered suitable since they differ considerably from the press releases in text genre, domain, and formal characteristics (e.g., the higher density of pronouns).

The five experiments that were conducted are: (1) deactivated syntactic parallelism, (2) deactivated syntactic subject salience $\wedge$ deactivated syntactic role hierarchy

magnitude since, in the key of the press release corpus, around 2,300 coreference classes are specified, the first textual mention of which is typically accomplished by a common noun or name.

41 Regarding the coreference class task, one must keep in mind that even the interannotator agreement that was measured during key construction in MUC-6 amounts only to 81%. Clearly, human performance with respect to the annotation of reference corpora, which, to a large extent, depends on the complexity of the task definition, imposes an upper bound on the system performance that is, in principle, measurable.

Table 3
Variation of preference strategies of ROSANA.

	News agency press releases corpus					
Experiment	P_{co}	R_{co}	P_{na}	R_{na}	**PER3**	**POS3**
ROSANA (orig.)	0.81	0.68	0.70	0.65	0.71	0.76
(1) −SYR	0.80	0.68	0.68	0.63	0.70	0.73
(2) −SUP, ...	0.80	0.68	0.67	0.62	0.69	0.73
(3) −SYR,−SUP, ...	0.78	0.66	0.58	0.54	0.56	0.69
(4) −SDM	0.78	0.66	0.60	0.56	0.63	0.60
(5) −KAM	0.80	0.68	0.66	0.61	0.65	0.77
	Mozart operas corpus					
Experiment	P_{co}	R_{co}	P_{na}	R_{na}	**PER3**	**POS3**
ROSANA (orig.)	0.88	0.81	0.75	0.74	0.79	0.77
(1) −SYR	0.89	0.82	0.76	0.75	0.77	0.80
(2) −SUP, ...	0.87	0.80	0.68	0.68	0.74	0.67
(3) −SYR,−SUP, ...	0.87	0.80	0.70	0.70	0.73	0.79
(4) −SDM	0.84	0.77	0.55	0.55	0.55	0.50
(5) −KAM	0.88	0.81	0.71	0.71	0.75	0.67

(salience of direct/indirect objects and adverbial PPs), (3) = (1) ∧ (2), (4) deactivated sentence recency, and (5) deactivated negative preference for cataphoric resumptions (cf. Section 4.5). The results are shown in Table 3, where rows correspond to the different experiments and columns[42] represent the most important evaluation measures.[43]

First, the table reveals that, with respect to the system's performance in interpreting *pronominal* anaphora, the model-theoretic scoring scheme must be regarded as an unsuitable indicator since the sensitivity with respect to the salience strategy variations is too low. Second, the results on the evaluation set of the press release corpus confirm the original assignment of salience factors and weights. Moreover, some interesting observations concerning the relative contributions of the factors can be made. The algorithmically trivial preference criterion of sentence recency proved to be the most valuable factor (Experiment (4)).[44] In the case of the press release corpus and regarding the PER3 measure, a relation of mutual substitution seems to hold between the factors of syntactic parallelism and subject salience/syntactic role hierarchy: whereas deactivation of either strategy results in a moderate performance reduction (Experiments (1) and (2), respectively), deactivation of *both* strategies induces considerable deterioration (Experiment (3)). These findings are in line with the results of Lappin and Leass (1994), who made similar observations in their factor variation experiments, but conjectured on a more abstract level that a relationship of complex *interdependency* holds between the different syntactic salience factors.[45] In providing evidence for a relation of *mutual*

42 In the columns labeled PER3 and POS3, the results in the immediate (*arbitrary*) antecedent discipline are shown.

43 The results of the original version of ROSANA on the Mozart operas corpus are: $(P_{oc}, R_{oc}) = (0.95,0.98)$, $(P_{co}, R_{co}) = (0.88,0.81)$, $(P_{na}, R_{na}) = (0.75,0.74)$; nonpronominal anchors for pronouns of type PER3/POS3 are determined with a precision of 0.70/0.76. Hence, performance is even better than on the press release corpus, a result that may be partly explained by the higher proportion of pronouns and names, which, as observed above, are resolvable with higher precision than definite NPs.

44 This result coincides with similar findings by numerous other researchers, for example, Lappin and Leass (1994, page 551).

45 Lappin and Leass took into account four groups of "structural" salience factors: parallelism,

substitution between two important classes of syntactic preference factors, the above results allow for a more precise rendering of this statement.

Closer investigation of the results on the Mozart operas corpus reveals that the factor assignment only partly generalizes. The result deterioration induced by deactivation of sentence recency (Experiment (4)) is even larger, a finding that may be explained by a characteristic property of the cohesion structure of opera plot texts, namely, the rapid shifts of the local foci from scene to scene, which contrasts sharply with the typically steady focus in the press release texts. If the sentence recency factor is switched off, this implies that the remaining focus-approximating salience factors (syntactic parallelism, subject salience, hierarchy of syntactic function) may lead to wrong decisions since the local foci of past scenes, which in the meantime have moved out of focus, would receive the same salience.

A similar observation of cohesion structure dependency may be made regarding syntactic parallelism in the POS3 strategy. Whereas, with respect to the press release corpus, the positive contribution of this factor is confirmed, for the Mozart opera texts, a negative contribution was measured: the deactivation in Experiment (1) yields a *gain* of 3%. Closer analysis of the documents reveals that local contexts contributing multiple POS3 occurrences with different reference are typical for this text genre:

(9) On a dark night in Seville, Leporello is keeping watch, grumbling, outside a house in which his master Don Giovanni is engaged in his latest amorous pursuit.

Again, the findings of the factor assignment experiments permit elaboration on a conjecture by Lappin and Leass (1994, page 552), according to which considerable improvement should be achieved by employing, for an optimization of the factor assignments, statistical analyses of patterns of pronominal anaphora in corpora. More precisely, the above results indicate that the *text genre* is reflected in some formal properties of the cohesion structure that are important clues for the choice of factors. In other words, the experiments indicate that salience factors and weights should be assigned in a *genre-specific* way.[46] From a practical point of view, these results are highly relevant since, meanwhile, various referentially annotated corpora of different genres (particularly the key data provided in formal evaluations) have been made available.

7. Comparison and Conclusion

In the previous sections, a robust approach to anaphor interpretation has been developed that follows the deficient description model. Based on a set of disjoint reference rule patterns for fragmentary syntax, the ROSANA system accomplishes coreference resolution with high precision and recall in various evaluation disciplines. The evaluation was carried out according to an enhanced set of scoring measures that sheds light on aspects of development as well as application. The different arguments put forward for an enhancement of the evaluation methodology for coreference resolution systems have been confirmed. In particular, the evaluation results have proven the

nonadverbial/matrix and head emphasis, hierarchy of syntactic roles, and cataphora penalty. Whereas, during individual deactivation of these factors, they observed comparatively small deteriorations of less than 4%, the combined deactivation led to a reduction of more than 25% (Lappin and Leass 1994, page 552).

46 In searching for additional evidence, further tests on corpora of other text genres should be carried out.

lack of sensitivity of the original model-theoretic coreference scoring scheme with respect to salience strategy variation as well as pronoun interpretation quality. Moreover, the figures confirm that, regarding pronouns, determining nonpronominal anchors is more difficult than computing arbitrary cospecifying antecedents: on average, results in the former discipline are 5.5% (press releases)/5% (Mozart operas corpus) below the results in the latter discipline.

As a proper basis for comparing ROSANA with the approaches of Lappin and Leass (1994) and Kennedy and Boguraev (1996), one must focus on the evaluation results for third person pronouns in the immediate (arbitrary) antecedent discipline (as discussed in Section 6.2). For third person pronouns (comprising nonpossessives, possessives, reflexives/reciprocals, and relative pronouns), ROSANA determined cospecifying antecedents with a precision of $\frac{322}{432} = 0.75$ (press releases) and $\frac{185}{233} = 0.79$ (Mozart operas corpus). At first sight, the gap between these figures and the precision of 0.86 that was determined for the (nonrobust) approach of Lappin and Leass is still considerable, a difference that may be partly attributed to the more difficult conditions of robust processing, and partly to the (presumably well-behaved) characteristics of the text corpus employed by Lappin and Leass (computer manuals). The standard of comparison, however, is the robust approach of Kennedy and Boguraev, which follows the shallow description model and which achieves an average precision of 0.75 on a broad set of texts taken from different genres and domains. In the case of ROSANA, a precision of 0.75 is achieved on the corpus of press releases, which exhibits the typical properties of mass texts (e.g., a comparatively high rate of orthographic and syntactic errors) and, hence, presumably imposes high demands on robust processing. On the Mozart operas corpus, which, in this sense, is easier, the scores are considerably higher. However, since different evaluation corpora have been used, and, moreover, since the precision figure mentioned in the results of Kennedy and Boguraev (1996) is not qualified with respect to text genre, a direct comparison of the empirical results should be based on further investigations.[47]

As an alternative way of comparing the two approaches, their performance with respect to the robust algorithmization of anaphor resolution strategies that rely on syntactic evidence (in particular, syntactic disjoint reference) may be evaluated in detail. Regarding the scope of robust processing of ROSANA according to the deficient description model, a qualification of the typical failures gives evidence that, with respect to the fragmentary descriptions generated by the chosen parser, the robust implementation of syntactic disjoint reference is nearly optimal. None of the 7 incorrect antecedent choices that are due to failures of the disjoint reference strategy (out of a total of 246 wrong antecedent choices for the evaluation corpus of press releases) are due to wrong predictions of the (still partly heuristic) algorithmization of the binding-theoretic restrictions; rather, they are caused by wrong (in contrast to fragmentary, i.e., partial) parsing results: while already employing defensive parsing strategies, the parser still overgenerates in certain cases. In 6 of the 7 disjoint reference failures, a configurationally admissible candidate has been erroneously eliminated; in the remaining case, a configurationally forbidden candidate has been erroneously approved. Hence, there is a tendency toward overgenerating disjoint reference restrictions. A detailed

47 This might be achieved either by employing the corpus used by Kennedy and Boguraev, which was not available at the time ROSANA was evaluated, or by running Kennedy and Boguraev's algorithm (a reimplementation of which requires, in particular, a formal description of the regular filter employed by Kennedy and Boguraev to partially infer constituency) on the news agency press releases and Mozart operas corpora.

analysis reveals that, in 4 of the 6 cases, the respective parsing error consists in a wrong interpretation of a structurally ambiguous relative clause. This gives evidence that, while the rate of disjoint reference failure is already very low (2.8% of all failures), a slight improvement may be achieved by employing a more defensive parsing strategy with a slightly higher level of syntactic fragmentation, which, by now, amounts to an average of 2.61 fragments per sentence.[48]

These findings may be compared with the results of Kennedy and Boguraev (1996), which report 2 cases of pronoun misinterpretation (out of a total of 75 failures) that are due to a failure to establish configurationally determined disjoint reference, and several additional cases in which a wrong antecedent was chosen because of a wrong heuristic assignment of syntactic salience factors. Hence, with respect to syntactic disjoint reference, the failure rate of the shallow description approach (2.7% of all failures) of Kennedy and Boguraev is of the same order of magnitude as that of the deficient description approach followed by ROSANA (2.8%). As the above analysis has revealed, however, in the case of ROSANA, the disjoint reference errors are not induced by failures of the heuristic algorithmization of the disjoint reference conditions; instead, they can be traced back to wrong decisions made during parsing, thus leaving some room for improvement by fine-tuning the rule pattern set and parsing strategy. This opens up the possibility of a further refinement, which is the immediate consequence of the conceptually transparent way of implementing robust disjoint reference by following the deficient description model. Regarding the syntactic preference strategies, ROSANA scores well, too: only 3 wrong antecedent decisions (1.2% of all failures) are due to errors in the assignment of syntactic salience factors.

Since the difference in interpretation quality can be expected to be small, the decision whether to follow the shallow description approach or the deficient description approach may be based on practical considerations. Whereas the former approach imposes lesser demands on preprocessing resources and implementation, the latter approach may yield slightly better results. If one considers implementation and fine-tuning of the deficient description algorithm relative to a particular parser as a once-and-for-all effort, and, moreover, takes into account the further benefits of having partial syntactic analyses available during anaphor resolution,[49] the deficient description approach may be the method of choice for robust anaphor resolution. Depending on the parser that is used and the characteristics of the texts to be interpreted, which, in large part, determine number and type of failures of robust syntactic disjoint reference, it may, in certain cases, be reasonable to apply a *hybrid strategy* that aims at avoiding, as far as possible, heuristic decisions: if the syntactic analysis yields sufficient evidence, deficient descriptions are employed; otherwise, shallow description rules are used.[50]

48 As emphasized in Section 3.3, since, in general, the choice of rule patterns for robust disjoint reference should depend on the parser that is used, an increase in the degree of parse fragmentation may give rise to extending the set of patterns. The general question of optimizing the choice of rule patterns, relative to a given parser, is an important issue that deserves further attention.

49 For example, Kennedy and Boguraev (1996) mention the problem of interpreting quoted speech separately from its surrounding context, a complex problem whose solution should be facilitated if richer syntactic information is available. In fact, ROSANA already employs several successful heuristics for interpreting anaphors in quoted speech, such as the handling of first person pronouns that occur in the subject position of a quoted sentence.

50 Since the number of disjoint reference failures is already low, the potential benefits of employing a hybrid strategy are limited. Whether there may be an additional contribution depends heavily on the degree to which the two strategies of robust syntactic disjoint reference differ with respect to their failure cases, which should thus be analyzed by an in-depth evaluation on large corpora.

Whereas the robust algorithmization of syntactic disjoint reference has thus been achieved in a nearly optimal way, with respect to the overall set of anaphor resolution strategies, there is considerably more room for improvement. Regarding the important case of pronoun interpretation, as determined in Section 6.2, more than 30% of the failures are due to the assignment of incorrect gender attributes, and another 30% are induced by cases that are beyond the horizon of the heuristic antecedent preference strategies. Under the conditions of robust, operational processing, whereas the former problem may be solved by a once-and-for-all improvement of the lexical resources, the latter case remains difficult since, in general, background knowledge will be needed. According to the results of the formal evaluation, at least a partial, genre-specific refinement of the preference strategies may be achieved in a manner compatible with the conditions of robust processing. While a more systematic investigation and evaluation of the latter issue is pending, the above results give rise to the expectation that, by exploiting the potential for further improvements, robust approaches to anaphor resolution should be able to achieve a precision of 0.8 (arbitrary antecedent discipline) and 0.75 (nonpronominal anchor discipline) even on difficult text genres like press releases.

Acknowledgments

The author is grateful to Pasi Tapanainen, who provided the parses for the texts of the evaluation corpora. Thanks also to the three anonymous reviewers for their helpful comments and suggestions on an earlier draft of this paper.

References

Carbonell, Jaime G. and Ralf D. Brown. 1988. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*, pages 96–101.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris Publications.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Hirschman, Lynette. 1998. MUC-7 coreference task definition, version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html

Hobbs, Jerry R. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Järvinen, Timo and Pasi Tapanainen. 1997. A dependency parser for English. Technical Report TR-1, Department of General Linguistics, University of Helsinki.

Karlsson, Fred, Atro Voutilainen, Juha Heikkila, and Arto Antilla. 1995. *Constraint Grammar: A Language-Independent System for Parsing Free Text*. Mouton de Gruyter.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Menzel, Wolfgang. 1995. Robust processing of natural language. In Ipke Wachsmut, Claus-Rainer Rollinger, and Wilfried Brauer, editors, *KI-95: Advances in Artificial Intelligence. 19th Annual German Conference on Artificial Intelligence*. Lecture Notes in Artificial Intelligence 981. Springer, pages 19–34.

Mitkov, Ruslan. 2001. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence* 15(3):253–276.

Stuckardt, Roland. 1996. Anaphor resolution and the scope of syntactic constraints. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 932–937.

Stuckardt, Roland. 1997. Resolving anaphoric references on deficient syntactic descriptions. In Ruslan Mitkov and Branimir Boguraev, editors, *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts*, pages 30–37.

Stuckardt, Roland. 2000. *Qualitative Inhaltsanalyse durch Computer - ein*

uneinlösbarer Anspruch? Untersuchungen zur algorithmischen Textinhaltserschließung am Beispiel der referentiellen Interpretation, Ph.D. thesis, Department of Social Sciences, Johann Wolfgang Goethe University, Frankfurt am Main. Also: Tenea-Verlag, Berlin.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1996. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, pages 45–52.