Coreferential objects in German: Experimental evidence on reflexivity

Sam Featherston, Eberhard-Karls-Universität Tübingen

While object coreference structures are possible in English, the equivalent German construction is marginal, and additionally seems not to obey Binding Condition B. Work on this thorny problem has been hindered by a lack of agreement about which structures are and are not grammatical, and by the unpredicted nature of some contrasts. We applied the methodology of magnitude estimation (Bard et al 1996) to produce replicable grammaticality data as a basis for further work and to test competing accounts. The data reveals that none of the previous accounts we consider is wholly adequate. We show that the phenomena can be economically accounted for by a range of violable grammatical constraints on surface structure such as the Law of Increasing Members together with a violable form Binding Condition B. We argue that the most adequate way of allowing these constraints to interact is in a constraint weighting model (Keller 2000), and show how such a model can be tightly constrained by deriving the constraint weightings directly from the experimentally obtained judgements.

Although there is considerable variation among the languages of the world in the use and reference of anaphors and pronomininals, there also seem to be surprising cross-linguistic similarities. The permissible bindings of anaphoric elements seem far more rigidly restricted than would be expected if they were merely motivated by factors such as pragmatics or interpretative strategies, and so researchers have seen binding phenomena as a rich source of data about the universals of human language. Experimentation has played a part in suggesting that pragmatic factors are not the sole determinants of pronoun reference. For example, Corbett & Chang (1983) used a sentence-end probe task to examine whether context would limit the antecedents accessed by a pronoun. Subjects were presented with a sentences with either a full NP or a pronoun, and then shown a probe word for them to make a lexical decision on ("Is this a real word?").

(1) Ellen aimed a pistol at Harriet, ... Probe Response time
   a. but Ellen did not pull the trigger. ELLEN fast
       HARRIET slow
   b. but she did not pull the trigger. ELLEN fast
       HARRIET fast

When the subject of the second clause was specified as Ellen as in (1)a, then the response to the probe ELLEN was faster than to HARRIET. When the subject was a pronoun that could refer to either Ellen or Harriet as in (1)b the response times were similar to the two probe words. Since the response time is shorter for more recently accessed items, this shows that the pronoun in (1)b reactivated both possible antecedents, even though one of them was a much more natural continuation of the sentence content than the other. Mere probability in context did not appear to constrain the mental reactivation of possible antecedents, therefore, which must be seen as supporting a grammar-driven model of pronoun reference and not a context-driven one.

There is however other linguistic data which is less easy to account for in grammatical terms. An example of this is coreference between a direct and an indirect object in German, which shows surprising features which are not well understood. While there have been a number of attempts to account for the data, these have been hampered by lack of clarity in the data set about what is and is not grammatical, as well as by the unusual nature of some of the apparent constraints in operation. In this study we employed the magnitude estimation methodology (Bard et al 1996, Cowart 1997, Keller et al 1998) in order to clarify this murky area. On the firm basis of this empirically obtained data, we are able to evaluate hypotheses relevant to this structure which have been advanced in the literature. Since none of them comes close to accounting for the full data set, we develop our own account of the phenomena, making use of cumulative and violable constraints (eg Keller 2000). Since our approach makes use only of assumptions motivated elsewhere in the literature and accounts for the previously unexplained German object coreference data, it offers strong support for this experimental method and constraint interaction model.

---

1 This work was carried out in subproject A3 Suboptimal Syntactic Structures of the SFB 441 Linguistic Data Structures, funded by the Deutsche Forschungsgemeinschaft. Thanks to Wolfgang Sternefeld, the project leader, and Frank Keller of Saarbrücken, for their ideas, suggestions, and support. The errors and failings remain my own.
Object coreference in German

Most reflexives in German are syncretic with other pronominals, but the third-person reflexive *sich* is distinctive, although it can have either dative or accusative case. Most frequently, it abides by Binding Condition A and denotes coreference with the subject, but there are some structures where coreference with an object seems possible, although restricted. Unfortunately, the precise nature of these restrictions is unclear, in part because the grammaticality judgements of the relevant examples are controversial. For example, Grewendorf (1984, 1985, 1988) attributes the restrictions on the binding of reflexives by objects to a noun phrase hierarchy of grammatical functions, arguing that the binder of a pair must always be higher up in the hierarchy than the bindee. The hierarchy he adopts is a fairly standard one encoding the intuitive notion of obliqueness, ie Subject < Direct Object < Indirect object < Instrumental < Adverbia<h>l < Genitive. This successfully captures some of the differences in binding possibilities, for it is clear that subjects are the preferred binders, and the more oblique grammatical functions are probably impossible as antecedents. The tripartite distinction between subject, objects, and the more oblique functions is thus empirically justified. However this hierarchy also necessitates that there be a clear grammaticality difference between a structure in which a direct object binds an indirect object and one in which an indirect object binds a direct object, other things being equal. Grewendorf argues that this is the case, and offers the following relevant judgements (2) (eg Grewendorf 1988).

\[(2)a\] Der Arzt zeigte den Patienten *sich/*ihm im Spiegel.
*the doctor showed the patient.DAT REFL/PRN.DAT in.the mirror*

\[(2)b\] Der Arzt zeigte dem Patienten *sich/*ihn im Spiegel.
*the doctor showed the patient.DAT REFL/PRN.ACC in.the mirror*

Example (2)a indicates that a dative reflexive may be bound by an accusative binder, but a dative pronominal cannot be. Example (2)b shows that an accusative reflexive cannot be bound by a dative antecedent, but that an accusative pronominal can. This account is superficially attractive since it links in to the noun phrase accessibility hierarchy which has been independently advanced for other purposes (eg Keenan & Comrie 1977, Pollard & Sag 1994). In addition, it captures the three-way split between subject, objects, and obliques quite well. However, there is a difficulty associated with it, namely that the data does not clearly support it. All object coreference structures in German are somewhat marked, which makes the judgements rather marginal, but even in this unclear data set other authors give very different judgements of sentences such as those in the set in (2). Next, it is phrased in terms of grammatical functions, but there is evidence that binding is also affected by structural factors such as c-command and surface factors such as linear precedence; it is unclear how these factors should interact in this model. This approach exists in other variants too. Primus (1987) too argues that the data is correctly captured by a hierarchy of grammatical functions, but adds to this a range of additional constraints, such as that a non-subject binder must c-command (the trace of) its bindee. This suggestion seems plausible, but is difficult to assess: the binding behaviour of complements is perhaps the major source of evidence we have for their c-command relations. Without some independent measure of c-command it is difficult to see how this proposal can generate any predictions additional to those of Grewendorf’s proposal. In this study we merely address the grammaticality of reflexives and pronouns in coreference structures, and explicitly test the prediction of Grewendorf that a direct object should bind an indirect object more readily than an indirect object binds a direct object.

Fanselow’s (1991) Proper Inclusion Principle suggests that a more specific rule will block the application of more a general one if its domain of application is properly included within it. The binding possibilities of reflexives and pronouns, he argues, are an example of this. Their distribution is not merely described by complementarity but actually driven by it: that is, the binding domain of the reflexive anaphor is not only co-extensive with the domain within which a pronoun must remain free, but also the definition of it. Notice that this approach has much in common with the pragmatic approach of Levinson (1991), in which a similar blocking mechanism applying to the maxims of felicity is argued to underlie binding conditions. Now while this complementarity-driven approach is readily compatible with Binding Conditions A and B, a number of exceptions to it have been noted, where the pronoun and reflexive both seem possible, for example in English:

(3) Kay and Tybalt pulled the blanket back over them/themselves

Fanselow does not discuss precisely our object coreference structures, but the prediction his theory makes is clear: where the included member of a pair of anaphoric elements can grammatically bind, the container of the pair must be ungrammatical. Now this is doubtful in object coreference structures in German; for example, Lechner (2000) describes (4) with a reflexive and (5) with a pronoun as “equally bad”.

(4) *weil ich die Maria sich im Spiegel gezeigt habe
*as I the M.ACC REFLL.DAT in.the mirror shown have

"as I have showed Maria herself in the mirror"
In our study we test whether the pronoun and reflexive are indeed in complementary distribution in these structures as predicted by the Binding Theory and the Proper Inclusion Principle.

Reis (1976) notes that these sentence types are much improved by the modification of the anaphor with a selbst ("self"). She raises the question whether this selbst insertion should be seen as a separate grammatical process, or whether it is merely a contextual factor; if it were the first, we might need to conclude that object binding of reflexives unmodified by selbst is actually ungrammatical in German. Reis herself suggests that examples without a modifying selbst which seem acceptable exist, eg (6)-(7).

(6) Hans überlässt die Schwester sich
"Hans leaves the sister to herself"
(7) Der Psychoanalytiker hat den Patienten wieder an sich gewöhnt
"The psychoanalyst has got the patient used to himself again"

Primus (1992) offers an account of the behaviour of selbst as a scalar focus adverb. While in most uses of selbst the scalar quality is visible, this is not salient in its use as modifier of a reflexive, but she argues that this is a difference of degree, and that it is not sufficient to force a separate categorization selbst as a reflexive. She points out that structures such as (8) are on the verge of ungrammaticality, except in a contrastive context in which Peter was in the habit of doing all these things to others but never to himself. Now if the primary content of selbst were emphatic reflexivity, this should not occur.

(8) Peter wäscht/kämmt/rasiert/setzt sich selbst
"Peter washes/combs/shaves/sits to himself"

On the other hand, selbst is fine in sentences such as (9), which show its focus function with no requirement for emphatic reflexivity. The situation consists of a simple distinction between family members; there is nothing remarkable about Maria preferring schnapps. There is however a contrast effect, and this requires a stressed anaphor. A bare sich is only marginally stressable, so the addition of the focus adverb is strongly preferred in such contexts. Now being the bearer of stress is readily compatible with a focus particle, and removes the need for selbst to separately bear reflexive semantic content. On this analysis therefore, there is nothing intrinsically reflexive about selbst, it merely collocates frequently with reflexives.

Reinhart and Reuland (ie Reinhart & Reuland (1993) and Reuland & Reinhart (1995)) offer another account of reflexivity, which has implications both for the distribution of anaphors and pronouns as well as for the nature of selbst. They distinguish between simplex SE-type and complex SELF-type anaphors (see also Pica 1984, Faltz 1985, and Everaert 1986), and argue that the former are in fact pronouns, not reflexives, but that they can occur as co-arguments where pronouns cannot because they are underspecified for phi-features, specifically gender, number and case, which makes them non-referential and thus feasible feet of chains. The disjoint distribution of SE- and SELF-type anaphors is achieved by a rewriting of Binding Condition B, which in its new form states that a semantically reflexive predicate must be reflexive-marked. This can happen in two ways: predicates which are grammatically reflexive are lexically reflexive-marked, other predicates must be reflexive-marked by having a SELF-anaphor as an argument. This analysis encounters a problem in German, since both lexically reflexive-marked verbs (10a) and verbs only optionally reflexive (10b) can have sich as an argument.
coreference structures. With verbs not lexically reflexive, only SELF-anaphors should be possible, since neither SE-
anaphors nor pronouns would reflexive-mark the predicate. This means that *sich* as an anaphor should be possible, with
or without a modifying *selbst*, while a pronoun should be impossible because it fails to reflexive-mark. Now the
authors do not discuss the possibility of a pronoun plus *selbst*, but the prediction is clear: if *selbst* is a focus adverb as
Primus (1992) claims (see above), a pronoun plus *selbst* should behave just as a lone pronoun - focus adverbs plainly
cannot reflexive-mark. If on the other hand *selbst* is an emphatic reflexive we might assume that it can reflexive-mark,
and the semantically but not lexically reflexive predicates in our object coreference structures should be no less
grammatical with a pronoun plus *selbst* than with a reflexive anaphor. Either way we should expect a clear pattern: a
pronoun plus *selbst* should either pattern just like a pronoun or just like a reflexive.

A final surprising feature of these constructions is noted by Elena Anagnostopoulou (cited in Lechner 2000). It seems
that pronominals are better than full NPs as antecedents in object coreference structures, so that (11) is more acceptable
than (12):

    *the hairdresser showed the customer himself in the mirror*

(12) ?Die Friseurin zeigte ihm, sich, im Spiegel.
    *the hairdresser showed him himself in the mirror*

This is of course somewhat surprising, since the NP type of the antecedent is not generally thought to play a role in
binding structures, and so we investigate this variable in our experiment too.

Let us note here that we shall not address reciprocal constructions. It seems to be generally agreed (eg Frey 1993,
Reuland & Reinhart 1995) that reciprocals are subject to additional constraints, and so we address only data with simple
coreference here. When we have an adequate account of the behaviour of reflexives we hope to investigate the special
case of reciprocals and discover what additional factors affect them.

Experimental design

Perhaps the major problem holding back progress in this area is the indeterminacy of the data. When most authors
contest some of the judgements in the preceding literature, it is clear that little theoretical advance can be made. In
particular it is sometimes maintained that there is in German no grammatical equivalent construction to the English (13).

(13) I showed Tanja herself in the mirror

Our first aim therefore is to determine whether object coreference is or is not grammatical in German in a range of
different forms whilst strictly controlling for factors such as lexis, plausibility, and context. There are a range of claims
in the literature about which examples are feasible, and we hope to provide answers based on hard data.

We further wish to test some of the accounts of this data which have appeared in the literature. The accounts which
make use of an obliqueness hierarchy predict that an object coreference structure in which a direct object binds an
indirect object will be better than the reverse. We test this in our experiment by including a parameter Case Order,
which has two values: Dat<Acc and Acc<Dat. We shall restrict ourselves to testing only structures in which the anaphor
precedes the antecedent in this experiment, since the reverse order, as for example in (14), is so marked that it would
be difficult to ensure that subjects evaluated the intended coreferential reading. For this reason, the Case Order
parameter effectively specifies the case of the binder and bound item.

(14) weil ich sich, den Paul, im Spiegel gezeigt habe
    *as I SELF the Paul in the mirror shown have*

Another factor we shall study is the effect of binding conditions A and B, since the data exhibits what Grewendorf
(1985) calls "a problem for GB Theory", namely the fact that the difference between reflexives and pronouns seems
not as clear-cut as we might expect. We therefore test each condition both with pronouns and with reflexives as
pronouns. Our next experimental parameter is that of the syntactic status of the binder. We are aware of no coherent
explanation why full NPs and pronominals should not bind equally well, and so we include it in our experiment in order
to see under which circumstances the effect holds. Our final area of interest in this construction is the role played by
*selbst*. We therefore test each condition both with and without *selbst*, and hope in this way to establish what role it is
playing. The experimental design thus consisted of four binary factors, resulting in sixteen different syntactic conditions,
presented in table 1.

The first column in the table contains the abbreviations for the syntactic conditions that we shall use here: the first letter
refers to the parameter Antecedent Type and indicates either an NP n or a pronoun p; the second indicates the Case
Order of the internal arguments, dative before accusative is \( d \) and accusative before dative is \( a \); the third shows the Anaphor Type, \( r \) for reflexive and \( p \) for pronoun; and the last denotes \( selbst \), an \( s \) showing that a \( selbst \) occurs, no \( s \) showing that there is none. Inspection of the codings and their realisations in the table should clarify this. Note that we shall also use an \( x \) in these codings to indicate "any value", so that the coding \( ndxx \) refers to all the sentence types which have a dative NP as the binder; similarly \( xxrs \) denotes all the conditions which terminate in \( sich selbst \).

<table>
<thead>
<tr>
<th>Code</th>
<th>Sentence form</th>
<th>Antecedent</th>
<th>Case Order</th>
<th>Anaphor</th>
<th>Selbst</th>
</tr>
</thead>
<tbody>
<tr>
<td>ndr</td>
<td>weil ich dem NP sich gezeigt habe</td>
<td>NP</td>
<td>dat &lt; acc</td>
<td>reflexive</td>
<td></td>
</tr>
<tr>
<td>ndrs</td>
<td>weil ich dem NP sich selbst gezeigt habe</td>
<td>NP</td>
<td>dat &lt; acc</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>ndp</td>
<td>weil ich dem NP ihm gezeigt habe</td>
<td>NP</td>
<td>dat &lt; acc</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>ndps</td>
<td>weil ich dem NP ihn selbst gezeigt habe</td>
<td>NP</td>
<td>dat &lt; acc</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>nar</td>
<td>weil ich den NP sich gezeigt habe</td>
<td>NP</td>
<td>acc &lt; dat</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>nars</td>
<td>weil ich den NP sich selbst gezeigt habe</td>
<td>NP</td>
<td>acc &lt; dat</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>nap</td>
<td>weil ich den NP ihm gezeigt habe</td>
<td>NP</td>
<td>acc &lt; dat</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>naps</td>
<td>weil ich den NP ihn selbst gezeigt habe</td>
<td>NP</td>
<td>acc &lt; dat</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>pdr</td>
<td>weil ich ihm sich gezeigt habe</td>
<td>pronoun</td>
<td>dat &lt; acc</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>pdrs</td>
<td>weil ich ihm sich selbst gezeigt habe</td>
<td>pronoun</td>
<td>dat &lt; acc</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>pdp</td>
<td>weil ich ihm ihn gezeigt habe</td>
<td>pronoun</td>
<td>dat &lt; acc</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>pdps</td>
<td>weil ich ihm ihn selbst gezeigt habe</td>
<td>pronoun</td>
<td>dat &lt; acc</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>par</td>
<td>weil ich ihn sich gezeigt habe</td>
<td>pronoun</td>
<td>acc &lt; dat</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>pars</td>
<td>weil ich ihn sich selbst gezeigt habe</td>
<td>pronoun</td>
<td>acc &lt; dat</td>
<td>reflexive</td>
<td>✓</td>
</tr>
<tr>
<td>pap</td>
<td>weil ich ihn ihm gezeigt habe</td>
<td>pronoun</td>
<td>acc &lt; dat</td>
<td>pronoun</td>
<td>✓</td>
</tr>
<tr>
<td>paps</td>
<td>weil ich ihn ihm selbst gezeigt habe</td>
<td>pronoun</td>
<td>acc &lt; dat</td>
<td>pronoun</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Experimental conditions

Experiment

The magnitude estimation methodology (Bard et al 1996, Cowart 1997, Keller 2000) allows us to obtain maximally differentiated judgements from a group of informants and compare them meaningfully. It derives from a methodology used to grade physical sensations such as heat and brightness and developed from there for use in attitude and opinion measurement (Stevens 1975). In order to remove the restrictions of the scale from informants’ judgements, it varies from standard elicitation of grammaticality judgements in several ways. The first innovation is that subjects are asked to provide purely relative judgements: at no point is an absolute criterion of grammaticality applied. Also, all judgements are proportional; ie subjects are asked to state how many times better or worse sentence A is than sentence B. Next, the subjects themselves fix the value of the reference item relative to which subsequent judgements are made. Furthermore, the scale along which judgements are made is open-ended: subjects decide its amplitude for themselves. Lastly, the scale has no minimum division: subjects can always produce an additional intermediate rating. The net result is that subjects are able to produce judgements which distinguish the differences they perceive with minimal interference from the scale.

Materials and Procedure

Eight basic texts were constructed which consisted of an experimental sentence with a double object binding construction and a context pragmatically motivating it, of the form of (15). Notice that the desired reading is forced by the context, and that this is reinforced by the explicit restatement of the intended reading in brackets.
The range of lexis and structures within these experimental materials was strictly controlled. The binding structure was put in an embedded clause, since this is generally regarded as the basic clause structure of German, and to our knowledge, no difference in binding between matrix and embedded clauses has ever been argued for. Five items used grammatically masculine gender nouns for the reflexive, three feminines. We utilized only NP-complements, although in the literature no difference is usually made between NPs and PP complements in this structure. In order to make a full NP and a pronoun equally good in the position of the binder, reference to the referent of the objects was excluded in the matrix clause. In addition, the subject of the embedded clause was always ich, to exclude the subject as a possible antecedent of the reflexive. We excluded verbs which have been experimentally shown to prefer acc-dat order of their complements (Mergel 1994), eg *vorspelen*. Since the verbs which semantically permit double object reflexives are few, we repeated the four which allow it most felicitously (*zeigen, vorspielen, zuweisen, empfehlen*). The homogeneity and plausibility of the materials were informally pre-tested in order to minimize variation between items.

Eight versions of these experimental materials were constructed such that each subject saw each syntactic condition once and each item twice, and each condition and item appeared equally often. The experimental materials of this experiment were mixed among 17 sentences which made up two other experiments so that the three mutually provided fillers. The total number of sentences each subject saw was 33. Subjects were recruited by flier in Tübingen university and by newsgroup listing, and were offered a financial incentive to take part. Twenty-seven subjects completed the experiment, but one subject’s data was excluded from further analysis because he failed to follow instructions. Participants were asked to supply their names, ages (mean age 27.3, range 17-46), sex (4 females, 22 males), occupations (16 identified themselves as students, most others by their faculty, eg “archaeologist”) and dialect backgrounds (17 from the southern areas of Baden-Württemberg and Bavaria, 6 from the central areas of Pfalz, Hessen and Nordrhein-Westfalen, 1 from the north, 2 claimed no dialect).

**Results**

The data was first normalized and then analysed using repeated measures anovas by Subject ($F_1$ and $p_1$) and by Item ($F_2$ and $p_2$). The normalization was achieved by subtracting the subject’s mean from each score and then dividing it by the subject’s standard deviation. The results are presented in figure 1, which shows the mean normalized grammaticality judgement score and 95% confidence interval for each experimental condition. Notice that higher scores indicate greater perceived grammaticality, but the relationships are only relative: there is no point which indicates absolute (un)grammaticality.

Both the graph and the statistical analysis show two very clear effects and two weaker ones. First a pronoun antecedent is better than an NP antecedent ($F_1=18.76, p_1<0.000; F_2=10.06, p_2=0.016$). To see this we compare the left-hand side of the graph (the *nxxx* conditions) with the right-hand side (p*xx*). Next, a *selbst* modifying the anaphor greatly improves the acceptability of the construction ($F_1=18.91, p_1<0.000; F_2=21.17, p_2=0.002$); this may be seen by comparing each *xxx0* condition with its equivalent *xxx* condition. It is generally true that the dative before accusative order is better than the inverse order (*xxdx* vs *xaxx*), but this is significant on the Subjects analysis only ($F_1=4.13, p_1=0.053; F_2=0.86, p_2=0.38$). Similarly, reflexives are on balance better as anaphors than pronouns (*ie xrx vs xxpx*), but this is significant only by Items ($F_1=0.35, p_1=0.56; F_2=6.97, p_2=0.033$). These latter two contrasts are less obvious because
there are some exceptions to them, but we believe these two to be partially obscured by another confounding factor.

Figure 1: Results of object coreference experiment; mean normalized judgements by structure.

To see this let us first examine the visual evidence of the graph. There are eight contrasts of reflexive anaphor versus pronominal anaphor (ie xxrx vs xxpx): six of them show a preference for the reflexive. This, we argue, is the background effect. Two other pairs, nar vs nap and par vs pap, show a clear reverse preference. Now consider the minimal pairs which make up the Case Order effect (ie xxvx vs xxxx). This contrast too is clear in six out of the eight cases, but the pairs ndp vs nap and pdp vs pap buck this trend. Lastly, let us look at the contrasting pairs with and without selbst (ie xxx vs xxxx). These too show a clear pattern for six of the eight contrasts, but which is contradicted by other two, here nap vs naps and pap vs paps. Now the interesting thing about these three pairs of exceptions to the general trends is that all three of them involve the structures nap and pap. Visual inspection of the graph confirms that exactly these two structures do not pattern with the others. The scores of each group of four structures form a slanted N shape, but just the nap and pap conditions break this pattern. They would both need to score lower to fit into the clear overall pattern. The visual impression of the graph is backed up by statistical facts: there is a statistical interaction of the factors of Case Order by Anaphor Type by Selbst Insertion ($F_1=4.84$, $p_1=0.037$; $F_2=2.23$, $p_2=0.179$), which shows that just these two conditions are behaving differently from the others. But what grammatical factor could improve the nap and pap conditions, without having a similar effect on the naps and paps conditions?

One factor which may be relevant is the reversal of the preferred case order of complements when these are both pronouns: while the usual unmarked order is dative before accusative, dative pronouns standardly follow accusative pronouns. This effect would cause the papx conditions to be better than otherwise expected, because they include the strings ihm ihm and sie ihr (ie PRN.ACC PRN.DAT in masculine and feminine forms). However, we should note that the equivalents with selbst are not affected, which would not be expected, if this were a effect of pronoun order.

We think that a different factor is causing this. The forms in question are in the subgroup which have a grammatical reading with non-coreferential objects. All conditions with a reflexive (ie sich) permit only the reading in which the antecedent and anaphor are coreferential; the same applies to all the conditions which have another marker of reflexivity ie selbst. None of these conditions permit a non-coreferential reading. There remain only the xpx conditions (ie ndp, nap, pdp and pap), in which the intended reading is enforced only by the context, and not by the grammar. We believe the feasibility in just these conditions of a non-coreferential reading to be the cause of the surprising improvement in the nap and pap conditions. Let us see how this might work: the subject reads . . . weil ich ihn ihm im Spiegel gezeigt habe and shudders inwardly at the poor acceptability of this structure with the coreferential reading and prepares to give it a bad score. But then the subject pauses and feels that the structure is surely not quite that bad. The string . . . weil ich ihn ihm . . . even sounds quite familiar. The subject upgrades the score s/he was planning to give and passes on.

---

2 One pair (pdrs vs pdps) shows only a marginal preference for the reflexive. This may be caused by the "ceiling effect": judgements at the upper margin sometimes show weaker differences that might have been expected.
The scores of the nap and pap conditions are better than expected therefore, because the intended reading is imposed only by the context sentence, and not by the grammatical form of the structure itself, and the sentence is feasible with the non-coreferential reading.

This account is thus restricted to the nap and pap conditions in two steps: firstly it can only apply to sentence types which are not grammatically constrained to a reflexive interpretation (ie those with neither sich nor selbst), secondly, it can only apply to those which are unacceptable with a coreferential reading, but acceptable with a non-coreferential reading. The four cases with neither selbst nor sich are ndp, nap, pdp, and pap. We do not see this effect with pdp because it is already quite acceptable as coreferential; we do not see it with ndp because it is unacceptable even as non-coreferential. This account is strongly supported by evidence from the time taken by subjects to input their judgements, which is automatically recorded by the experimental set-up. The nap and pap conditions were by far the slowest to be responded to (nap 17391ms, pap 18917ms, mean of others 14535ms, of which the fastest was paps 15994ms). This is strong evidence for our hypothesis that the non-coreferential reading is causing subjects to reconsider and upgrade these conditions in line with their felicity in a non-coreferential reading.

To summarize: it seems fairly clear that the weak statistical significance of the Case Order and Anaphor Type parameters is due to the quirky effect in the nap and pap conditions, which in turn is probably caused by the feasibility and greater naturalness of the non-coreferential reading of the two arguments in these conditions. We conclude therefore that all our four experimental parameters produced clear differences.

Testing the accounts

In this section we shall evaluate the attempted linguistic accounts of object coreference that we introduced in our syntactic sketch of the construction above. We first address the claim of Grewendorf (1984, 1985, 1988) and Primus (1987) that a hierarchy of grammatical functions is involved in object coreference structures, specifically that a binder must precede its bound item on the obliqueness hierarchy Subject < Direct Object < Indirect object < Instrumental < Adverbial < Genitive. We have found no support for such a model, as they clearly predict that binding of an indirect object by a direct object will be better than binding of a direct object by an indirect object. In fact in our experimental data the opposite is true: the structures with indirect objects binding direct objects are consistently better than the reverse. Since our experimental structures are syntactically and in part lexically identical to the examples used in the literature, we may be sure that our findings are relevant to the phenomenon. It seems reasonable to conclude that the obliqueness hierarchy is at least not the major determinant of permissible binding relations.

We next turn to the issue of complementarity of binding domains. A range of approaches including Levinson’s (1991) pragmatic reduction of the binding theory and Fanselow’s (1991) Proper Inclusion Principle claim that Binding Conditions A and B may be reduced to a single constraint, and that the distributions of pronouns and anaphors are mutually exclusive. These approaches thus predict that, all other things being equal, the acceptability of pronouns and reflexives will be inversely proportional: to the extent that one of the pair is possible, the other should be impossible. Our results do not support this, however; the structures with bound pronouns are consistently weaker than the reflexives, but the difference is not large, and there is no sign of complementary distribution. The appropriate generalisation seems to be that where a reflexive is possible, a pronoun is too, but slightly less so. The Proper Inclusion Principle as a determiner of binding domains is not, on the face of it, compatible with this data, and so its role in determining binding domains must be questionable.

It is worth noting however that this result does not actually exclude the possibility that the binding domains of reflexives and pronouns are complementary. The examples of binding condition satisfaction and violation usually advanced reveal a strong contrast between reflexives and pronouns, as in (16).

\[(16) \quad \text{a. Narcissus admired himself,}\]\n\[\text{b. *Narcissus admired herself,}\]\n
But these examples are standardly of binding by the subject. If, as seems likely, binding by any non-subject antecedent incurs a violation cost (cf Wexler & Manzini 1987), it may be that the clear grammaticality degradation we perceive in (16) is due to the violation of two constraints: Binding Condition B and the constraint which specifies that subject binders are preferred. In our experimental materials the binders are never subjects, and the bound pronouns produce a much weaker violation effect. The mismatch between the standard examples such as (16) and our results might thus

\[\text{3 To test our view that the nap structure is possible and the ndp structure is not, we carried out searches in the COSMAS database of German (Institut für Deutsche Sprache, Mannheim). The string den NP ihm occurs 7 times, but dem NP ihm not at all. So it would appear that the nap condition is a feasible structure of German while the ndp is not.}\]
be caused by the failure of the standard examples to factor out the portion of the grammaticality degradation due to the constraint specifying subject binders. We conclude therefore that our results probably require some revision to the standard version of the binding theory, but this need not necessarily involve the abandonment of the complementarity of the binding domains, only the downgrading of its violation cost. We hope to test this in further work.

The third syntactic analysis we considered was Reinhart & Reuland’s (1993) account of reflexivity in terms of a distinction between simplex SE-type and complex SELF-type anaphors and their rewrite of the binding conditions. The lack of complementarity between reflexives and pronouns must come as a surprise to this approach to reflexivity too, but the implications are different. On this account, the contrast is between a predicate used reflexively which is reflexive marked (by a SELF-anaphor) and one not reflexive-marked (by a pronoun). The relatively weak degradation for pronouns in this position must come as a particular surprise, since these violate two constraints, namely, Reinhart and Reuland’s binding condition B (since the reflexive predicate is not reflexive-marked) and their chain condition (since pronouns are +R(eferential)). One response that Reinhart and Reuland might give to this data would be for them to analyze selbst as an emphatic reflexive (contra Primus 1992), which could then credibly be argued to reflexive-mark independently. This would explain why the pronoun plus selbst combination is so much better than the bare pronoun, but these structures still fall foul of the chain condition, and so they should still be judged ungrammatical. In fact a pronoun plus selbst is judged better than a bare reflexive (sich), for which this account offers no explanation. This account would also not predict that the reflexive plus sich should be so much better than a bare reflexive. If sich is a SELF-anaphor then it can reflexive-mark; additional reflexive-marking should not be necessary. We may perhaps summarize that this approach is not supported by this data set.

One final issue which our results throw light on is the status of selbst. We noted above that there is controversy about whether selbst is merely a focus adverb or whether it has independent reflexive content. In our results, the effect of selbst is consistent across conditions, which might be seen as supporting Primus’ (1992) view of it as a focus adverb. A bare sich or ihn can only marginally carry focus and therefore stress, but this can occur very naturally with a modifying selbst. If selbst were a reflexive we should expect it to have a greater effect on the pronouns than on the reflexives: it should improve the pronoun scores (xxp vs xbps) more than the reflexives (xxr vs xxr) by adding reflexive content to the pronouns, which they lack. There is no apparent differential effect between the reflexives and the pronouns, however, suggesting that the effect of the addition of a selbst is the same whether the bound item is a reflexive or a pronoun. This would seem to support the interpretation of selbst as a focus adverb.

Linguistic constraints

Since none of the linguistic accounts of object coreference we tested was able to account for our experimental findings fully, we shall examine what sort of account might do so. To do this it is first necessary to consider how we might interpret our experimental variables as linguistic constraints. In a subsequent section we shall discuss how these constraints interact, but let us make it clear that we understand all these constraints as viable, that is, we accept that grammatical counter-examples may occur; they may thus be thought of more as descriptive generalisations than hard rules. For example, a reviewer comments that sentences such as (18) might be seen as problematic for our Binding Direction Rule below, which states that an antecedent must linearly precede the item it binds.

(18) als sich der Mann im Spiegel sah
    when REFL the man in.the mirror saw

Now on a standard view of grammatical constraints as inducing ungrammaticality when violated, this would naturally be true, since the reflexive precedes its antecedent, but this is not the assumption we make here. In example (18) two constraints or generalisations would be expected to apply, one of which reflects that fact that antecedents linearly precede their bound items, but the other one constrains pronominals to precede full NPs in the Mittelfeld. This example such as (18) does not show that the former does not hold as a descriptive generalization, merely that it is violable, by which we mean that the structure is not necessarily ungrammatical if it is violated. But even if violable, both of these constraints are syntactic regularities worthy of study.

Two linguistic constraints can be derived straight from the parameters of the experimental design. The parameter Anaphor Type showed that reflexives were better than pronouns as bound items. Since this is precisely the content of Binding Condition B, we adopt it here in a standard form.

(19) BC-B: Pronoun binding condition
    A pronoun must be free in its binding domain

Let us note two things: first, the violation cost of BC-B may be rather less than generally assumed, due, as we noted above, to a confound in the standard examples between the requirement for a reflexive to be bound and the preference for binding by a subject. Next, this condition must be recognized to be violable - its violation does not necessarily
reduce a sentence to gibberish. With the addition of these two provisos, we can readily accept that Binding Condition B applies but also that a structure with a bound pronoun such as weil ich ihm ihn selbst im Spiegel gezeigt habe ("because I showed him.DAT him.ACC SELF in the mirror") is nevertheless rated the second best variant we tested, only a little worse than the best variant, the equivalent structure with a reflexive.

The experimental parameter Case Order reveals that those structures with the dative preceding the accusative were judged better. This sort of preference has often been attributed in the literature to a case-related binding preference, a specific case of which is the obliqueness hierarchy. Our experimental result must make it very questionable whether such a hierarchy of obliqueness plays any role in binding conditions, in spite of some claims in the literature (eg Pollard & Sag 1994). Admittedly, the result does not exclude a role for a hierarchy of binders, but it does place strict limits on it. The standardly obliqueness hierarchy with direct objects dominating indirect objects is for example not possible, but a hierarchy in which objects, direct and indirect, occupied an equal position, on the other hand, would be possible, as would one in which indirect objects dominated direct objects. We doubt any such effect, however, since we can derive the same empirical effects more economically.

In fact there seems no obvious reason to attribute this preference to a facet of binding at all, but rather to a linear precedence restriction, which we may formulate as the Case Alignment Rule

(20) Dat<Acc: Case Alignment Rule
In the mittelfeld a dative precedes an accusative, unless they are both personal pronouns.

This too is a rule of German so familiar that it can be called a descriptive generalisation (Behaghel 1932, Lenerz 1977, Uszkoreit 1987, Pechmann et al 1994). We shall not investigate the cause of this effect here, but note that we have observed its effect in other experimental studies with non-coreferent objects. There seems little reason to stipulate a binding-specific hierarchy of grammatical functions, when the same explanatory work can be done by a well-attested word order preference.

The parameter Antecedent Type showed that pronominal binders were better than full NP binders. This has been regarded as puzzling (eg Lechner 2000), but it is readily explicable once one realizes that this too is not a constraint on binding but on constituent ordering. This result just shows that putting a full NP in front of a pro-form incurs a violation cost, which holds quite generally, even if they are not co-referential, see (21).

(21) a. weil ich dem Peter es gezeigt habe
   because I the.DAT Peter it shown have
b. weil ich es dem Peter gezeigt habe
   because I it the.DAT Peter shown have

This effect is frequently attributed to the Law of Increasing Members (Behaghel 1909, Lenerz 1977), which states that longer constituents should follow shorter ones. Since the binder precedes the bound item, and the bound item is necessarily a pro-form of some kind so that it can have anaphoric reference (Binding Condition C), it follows that the binder must be a light pronominal element too. We thus need make no new stipulations to derive this result, since it is merely an interaction of a constraint of the direction of binding (22) (eg Barss & Lasnik 1986, Jackendoff 1990), Binding Condition C (23) and the Law of Increasing Members (24).

(22) BI<BO: Binding Direction Rule
A binder must linearly precede its bound element.

(23) BC-C: Binding Condition C
A referring expression must be free.

(24) LIM: Law of Increasing Members
Heavier constituents follow lighter ones.

Notice that this approach can also deal with one other slightly puzzling feature of our results: the effect of the addition of a selbst does not appear to be consistent. The difference between the ndxx conditions with selbst and without selbst is larger than the equivalent difference between the pdxx conditions. It is a natural assumption that the addition of selbst to a pro-form increases the weight of the constituent, and thus the status of the sentence under the Law of Increasing Members (LIM); an NP preceding a bare pro-form is a clear violation, but an NP preceding a pro-form plus selbst is only a partial violation, since the greater length makes the weight balance more even. It seems, nevertheless, to be reasonable to expect some LIM effect here since, first, the string consists of a full NP preceding a pronominal, which is the marked order (eg Lenerz 1977), and second, the NP consists of three syllables (eg dem Martin), the pronominal phrase only two (eg ihm selbst). On this account, therefore, the pure effect of selbst insertion is of the magnitude seen in the pdxx conditions; the large differences between the ndxx types with and without selbst are the cumulative effect of the selbst insertion and the reduction in the LIM violation cost. Since this account in terms of the LIM neatly
accounts for both the puzzling features of the data, the variation in the Selbst Insertion effect as well as the Antecedent Type effect, it must be regarded as well-motivated.

The least straightforward effect to account for in syntactic terms is that of the addition of selbst. We noted above that the pattern of data in our result provides prima facie support for Primus’ (1992) analysis of selbst as a focus adverb with no reflexive content. In fact, this result provides suggestive evidence that the commonality between focus and reflexivity may go further than Primus suggests. Primus’s position entails that selbst can add no reflexive content to the pronoun since it is purely a focus particle. This is possible, but this seems to us to be more strongly undermined than Primus admits by data from compounds where the selbst does seem to carry precisely reflexive content (e.g. selbstkritisch “self-critical”, selbstmord “suicide”, selbstbewusst “self-aware”) which contrast with others where it has a merely emphatic content (e.g. selbständig “automatically”, Sichsthetigung “own contribution”, Selbstfinanzierung “self-financing”). The alternative would be that selbst has two separate functions, as a focus particle and as a marker of reflexivity. This is however also unattractive, as it is theoretically inelegant and because we detect no sign of differential effects in our results. If selbst were able to function as a reflexivity marker, we should expect it to improve the scores of our pronouns (which lack a reflexive semantic component) more than the reflexives, but this does not occur. Neither of the standard views succeeds in fully accounting for the facts, therefore.

The third possibility is more speculative, but it is at least consistent with the full set of evidence so far. It seems plausible that the key feature which distinguishes a reflexive from a standard pronoun is the ability to bear focus and its overt reflex, stress. Rather than just analysing selbst as a focus marker therefore, we are suggesting that the core of reflexivizability is focusability. Although this conclusion is novel, quite a lot of evidence seems to converge in support of it. It is firstly the simplest explanation of our own empirical results. Both reflexives and pronouns are judged better with selbst because it allows them more easily to carry focus. There is no differential effect on the pronoun because the factor that makes any pro-form usable as a reflexive is precisely the ability to be focused. This account is compatible too with the observation that lexically reflexive predicates resist the addition of a selbst (Primus 1992). It seems clear that the simple reflexive marker sich is the unstressed form, appropriate in cases where the reflexivity is lexically determined, such as (25). In cases of semantic but non-lexical reflexivity, the stressed form is necessary (26). Note that this distinction probably underlies the difference between Reinhart and Reuland’s (1993) SE-type and SELF-type anaphors too: the former bear no focus or stress, and the latter do. The fact that this distinction largely falls together with the distinction between predicates which are and are not lexically reflexive is hardly surprising — lexically reflexive predicates do not require any focus on their reflexives.

(25)   Heinrich benimmt sich
       Heinrich behaves REFL
(26)   Heinrich betrügt sich ?(selbst)
       Heinrich cheats REFL

Now these points would not alone support the hypothesis that reflexivizability reduces largely to focusability, but there is other data which provides a far firmer underpinning. First among these is the ability of perhaps any nominal, full NP or pronominal, to be used as a reflexive if stressed. For example, there exist at least three books called I like me (ISBN 0-307-16187-0, ISBN: 0-960-85161-5 and ISBN: 0-140-50819-8), a reading scheme for school children in the US called the same thing, and a Scottish punk band. Full NPs can be used reflexively too if given contrastive stress: Peter doesn’t love Janet, Peter loves PETER). These simple facts otherwise unaccounted for fall into place if the essential ingredient in reflexivizability is the ability to bear focus. Notice also that this offers an account of a puzzling fact noted by Dowty (1980), namely that first person pronouns are more easily used as reflexives than third person pronouns. Further, this hypothesis offers some account of why cross-linguistically reflexives and emphatic pronouns often have the same forms. Now let us not overstate this; there are also differences between focus and reflexivity — some reflexives are clearly grammaticalized as such and do not resemble emphatic pronouns. But this does make some sense of our finding that pronouns are only a little worse than reflexives as anaphors in double object coreference structures, and of the basic linguistic fact that most nominals can function reflexively if stressed.

In our experimental sentences, none of the verbs used were lexically reflexive, and so they required a stressable reflexive. Those conditions without selbst thus bore a violation cost. So it seems that the basis of the selbst insertion effect in our results is the need for a focusable (and therefore stressable) constituent in a syntactic position bearing focus. Now this is by no means specific to reflexive structures, an unstressed pronoun in a position which demands a stressed one can cause severe ungrammaticality, and not just stylistic oddness, as we can see in examples such as (27). A topicalised object needs to be stressed. The den pronoun type accepts this stress much more easily than an ihn pronoun type. The contrast between ihn and den in (27) seems parallel to the contrast between sich and sich selbst in (26). We therefore phrase the relevant constraint as the Stress Faithfulness Constraint in (28).

(27)   ?ihn/den hat Heinrich nie angelogen
violations which contribution to the winnowing process are applied. It is not clear that there is any empirically accessible difference between these two possibilities.

The grammar has more in common with the selection function should be thought of as operating probabilistically. The Decathlon Model contains an element of the highest scoring candidate will be selected, other high-scoring candidates can be selected for output, since the candidates to choose from. It is at this stage that the competition occurs, unlike in OT. Although in the default case selected by default. Selection in the Decathlon Model is more complex since the selection function still has multiple gradable evaluation scale, the values of which represent an interaction of the number and gravity of constraints violated. In evaluation OT uses merely a binary scale: candidates are either excluded or not yet excluded: competition of grammaticality, not a binary distinction (Keller 2000). The Decathlon Model is often assumed to be a variant of the OT approach, and indeed it does utilize a competition function (eg Last Resort) but the default assumption within generative grammar has been that anything is grammatical which does not violate a constraint, and nothing is grammatical which does.

At the other end of the scale we have Optimality Theory (OT)(Prince & Smolensky 1993). This is based upon the concept of competition, since grammaticality is not an absolute quality of a linguistic structure, but a feature of a relationship between a set of structures. In OT, constraints are applied sequentially in a ranking order. After the application of each constraint, the candidates which violated it are excluded, and the next constraint is applied: this process iterates until only one candidate remains. This last candidate thus optimal, and the assumption is that this one candidate is grammatical and will appear in the output. This interaction type has two important differences from the "live wire" model above: firstly, it requires that constraints be ordered in their application – this is the basis of the OT constraint ranking – and second, it requires a "smart" constraint application function: constraints are not just applied blindly and across the board as in the live wire model; only those constraints are applied which will contribute towards selecting a single candidate as output; those which will not distinguish between remaining candidates are not applied. It is a basic assumption of this approach that we can expect one and only one grammatical output for each underlying structure, and we should expect a clear difference between the output and all other imaginable realizations, but no differences between these unsuccessful candidates.

The constraint interaction we adopt here (following Keller 2000, but see also Uszkoreit 1987, Jacobs 1988, Pafel 1991, Gibson & Broihier 1998) we shall term the Decathlon Model. Decathletes take part in ten different subdisciplines and their performance in each is converted into numerical form and summed, the winner being the contestant with the highest overall score. The crucial linguistic fact about this model is that a constraint violation is not assumed to be immediately fatal, but rather simply to incur a violation cost. These costs are cumulative, which predicts a continuum of grammaticality, not a binary distinction (Keller 2000). The Decathlon Model is often assumed to be a variant of the OT approach, and indeed it does utilize a competition function. In both OT and in the Decathlon Model structures compete and normally one structure will end up best and appear in the output. However, if we look at the processes more closely, their differences emerge; to do this we must distinguish between the evaluation and selection functions involved. In evaluation OT uses merely a binary scale: candidates are either excluded or not yet excluded: competition in OT takes place within the evaluation function. The Decathlon Model on the other hand allows a potentially infinitely gradable evaluation scale, the values of which represent an interaction of the number and gravity of constraints violated.

OT’s selection procedure is trivial, since the evaluation process continues until exactly one candidate remains: it is selected by default. Selection in the Decathlon Model is more complex since the selection function still has multiple candidates to choose from. It is at this stage that the competition occurs, unlike in OT. Although in the default case the highest scoring candidate will be selected, other high-scoring candidates can be selected for output, since the selection function should be thought of as operating probabilistically. The Decathlon Model contains an element of competition, but this is in the output selection, not in the grammar itself. The grammar has more in common with the

\[ \text{Him has Heinrich never lied to.} \]

\[(28) \text{STR-F: Stress faithfulness constraint} \]

The STRESS feature on a syntactic position and the constituent occupying it must match.

In this way we can answer Reis’s (1976) question about the nature of the selbst insertion effect using only rather natural assumptions about focus and stress.

To sum up, the interaction of our four constraints STR-F, LIM, DT<AC and BC-B gives us the very satisfactory situation in which we are able to account for the data with only constraints which are motivated in the grammar more generally and no additional stipulations about binding at all.

**Constraint interaction**

The question now arises how the linguistic constraints which we have derived should interact. Perhaps the standard assumption within generative grammars is that any structure which breaks a rule is ungrammatical; this is thus a "live wire" model - any infringement is fatal. All rules are seen as independent, and any structure must abide by all of them to be grammatical. In its pure form this model permits multiple grammatical surface structures derived from identical underlying structures as well as underlying grammatical structures for which there is no grammatical surface form. In practice, generative syntacticians have recognized that there may be some variation in violation costs between rules (eg ECP vs subadjacency), and introduced elements of competition (eg Last Resort) but the default assumption within generative grammar has been that anything is grammatical which does not violate a constraint, and nothing is grammatical which does.

An alternative would be a "smart" violation cost application function: each constraint is applied in turn but only those violations which contribution to the winnowing process are applied. It is not clear that there is any empirically accessible difference between these two possibilities.
live wire interaction type: like this and unlike OT it applies all constraints to all structures blindly and unordered, and it uses an absolute, not relative conception of grammaticality.

Our empirical data gives us little choice in interaction type. It shows a continuum of grammaticality: there is no single clearly optimal candidate; our best two candidates are almost identical in their scores, and they are only slightly better than several others. It also reveals clear evidence of violation cost cumulativity: any structure with Acc<Dat word order is worse than its equivalent structure with Dat<Acc word order, and the difference between them is quite consistent. There is no sign in this data of stronger constraints blocking the effect of weaker constraints. These two features are plainly inconsistent with the pure versions of OT and the live wire model. Now, versions of OT which have been modified to permit cumulative effects do exist (eg Fischer 1999, Keller 2000) and may be able to deal with this type of data, but only at the cost of moving some considerable way from classic OT. Similarly, it is not actually inconsistent with generative grammar to assume a quantifiable and cumulative violation cost for each constraint type, it is merely not the standard assumption. Nor are the three models we mentioned above the only possibilities: a range of hybrids can be readily imagined.

The various models have both advantages and disadvantages, but we do not believe that any model yet captures the full complexity of constraint interaction. For example, while there is strong evidence that multiple constraint costs are cumulative in the area of marginal grammaticality (Keller 2000), our own work would tend to suggest that towards the top and the bottom ends of the grammaticality scale, the cumulativity reduces sharply: multiple constraint costs seem therefore follow a sigma function pattern. This might be accounted for by positing that our perception of grammaticality is sensitive in the marginal area, but becomes less acute towards the ends of the scale. Speakers’ concepts of "fully grammatical" and "fully ungrammatical" are on this view just an epiphenomenon of diminishing discrimination towards the boundaries; the ceilings and floors for judgements apparent, not real. Space does not allow us to discuss the requirements of a more adequate model in detail here and so we shall simply adopt an informal linearly additive constraint weighting model (compare Keller 2000), pending further work.

Testing the model

We have so far specified two variables in the model: the constraints involved in producing our empirical results and the way that these will interact. But we still need to fix one more parameter in order to have a functioning model, the violation costs of the constraints. We can then test our hypotheses by determining whether the summed weights of the constraints violated by each sentence type produce a pattern of data similar to our experimental results. If the observed results and the predicted results of our model are similar, then our approach will be supported.

In order to maximise the empirical base of our model, we take our constraint weightings directly from the experimental data. The constraint weight is the mean observed severity of each violation type, obtained by subtracting the normalized mean score for the conditions violating a constraint from the equivalent conditions fulfilling it. Notice that we disregarded the data from the nap and pap conditions in these calculations, since we have strong evidence that these are tainted with an extraneous factor. STR-F has a violation cost of 0.57, DT<AC 0.31, and BC-B 0.23. As we argued above, LIM seems to have a scalable violation cost, with strings consisting of NP+proform worse than NP+proform+selbst. We therefore calculate two separate weights: the full violation cost of 0.50 for NP+proform, but the partial violation cost of 0.30 for NP+proform+selbst. These calculations give us the constraint weightings in table 1. Notice that the Case Alignment constraint (20) contains an exception for pronoun pairs (Acc<Dat is the default order for two pronouns) and so the pap condition does not violate it.
On the left of the table, we see the sentence types and their codes. The central columns of the table show the violation costs incurred by each sentence type from each constraint, and the column Total Violation Cost sums these. The rightmost pair of columns show our experimental results, first the normalized experimental score of each sentence type and then its position in the experimental ranking. We order the conditions in the table by this ranking, since this represents the data that our constraint model must account for. We include the dubious nap and pap conditions here, but expect them to diverge somewhat. If our constraint model is successful, we should expect a correlation between the predicted scores in the Total Violation Cost column and the experimental results in the final two columns. In fact the match is good, the only major differences between the experimental order and the predicted order being nap and pap. Figure 2 reveals the correspondence visually. Here the cumulative violation costs are represented as negatives: this is merely a notational change to facilitate visual comparison.

<table>
<thead>
<tr>
<th>Code</th>
<th>Sentence form</th>
<th>Constraint and violation cost</th>
<th>Total violation cost</th>
<th>Experiment score</th>
<th>pos.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>STR-F</td>
<td>LIM</td>
<td>DT&lt;AC 0.31</td>
<td>BC-B</td>
</tr>
<tr>
<td>pdrs</td>
<td>ihm sich selbst</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pdps</td>
<td>ihm ihm selbst</td>
<td>0.3</td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>ndrs</td>
<td>dem NP sich selbst</td>
<td>0.3</td>
<td></td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>pars</td>
<td>ihn sich selbst</td>
<td>0.31</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>pdr</td>
<td>ihm sich</td>
<td>0.57</td>
<td></td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>pap</td>
<td>ihr ihm</td>
<td>0.57</td>
<td>Ø</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>ndps</td>
<td>dem NP ihm selbst</td>
<td>0.3</td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>nars</td>
<td>den NP sich selbst</td>
<td>0.3</td>
<td>0.31</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>paps</td>
<td>ihn ihm selbst</td>
<td>0.31</td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>pdp</td>
<td>ihm ihn</td>
<td>0.57</td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>nap</td>
<td>den NP ihm</td>
<td>0.57</td>
<td>0.5</td>
<td>0.31</td>
<td>0.23</td>
</tr>
<tr>
<td>naps</td>
<td>den NP ihm selbst</td>
<td>0.3</td>
<td>0.31</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>par</td>
<td>ihn sich</td>
<td>0.57</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>ndr</td>
<td>dem NP sich</td>
<td>0.57</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nar</td>
<td>den NP sich</td>
<td>0.57</td>
<td>0.5</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>ndp</td>
<td>dem NP ihn</td>
<td>0.57</td>
<td>0.5</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The correlation of experimental results and our predicted violation costs
To confirm the clear covariance of the two measures we performed statistical tests of the correlation. The Pearson’s correlation coefficient is 0.869 (\( p < 0.000 \), two-tailed) and Spearman’s =.910 (\( p < 0.000 \), two-tailed), even including the tainted nap and pap conditions. To conclude: our linguistic factors and experimentally obtained weighting show a remarkable degree of correspondence.

**Conclusions**

Many devices have been posited in the literature on Binding Theory to account for awkward data, but for the structure we investigated, object coreference in German, we have shown that no theoretical devices such as reconstruction, trace deletion, and accessibility hierarchies are necessary to account for the recalcitrant phenomena. An empirically adequate and theoretically economical account can be developed using only conditions on well-formed structures amply motivated from elsewhere in the literature and overtly visible in the data: we do not need to stipulate abstract and unverifiable mechanisms. But not only are the constraints we utilize well founded, so too is the body of experimentally obtained and replicable judgements on which we base them. Our empirical approach also permits us to take a more sophisticated view of grammaticality judgements. Since intuitions of our subjects about sentence grammaticality take the form of a continuum, we allow this representation of them, rather than squeezing them into an binary opposition of good and bad for theoretical convenience. This has not only the advantage of providing more detail about perceived grammaticality, but also permits us to claim that we are developing a more psychologically real grammar, in that it employs one abstraction from the data fewer (Schütze 1996, Cowart 1997).

Let us recognize that an essential part in our approach is the adoption of the Decathlon model of cumulative constraint weightings, in which all constraints apply blindly and exceptionlessly, and violation costs are summed. It will be clear that neither the familiar live wire model of constraint interaction used in generative grammar, in which any broken rule necessitates ungrammaticality, nor the OT approach, which requires the ordered application of constraints and the stipulation that only one output can be grammatical, can either deal with our data or produce the same results.

It is frequently objected that constraint weighting or ranking mechanisms are unconstrained, because the linguist has complete freedom to rank constraints and some degree of freedom to choose which constraints which will allow the description to capture the data set. Since constraint rankings do not obviously correspond to anything in the real world and are generally applied only to one data set, not generalized, these accounts suffer from untestability and lack of predictive capacity, it is argued. These criticisms do seem to have some element of truth particularly when applied to OT, where the two variables of constraint choice and constraint ranking may each be manipulated to produce the result.
the linguist requires.

Our application of experimental methodology makes us less vulnerable to these criticisms. First, our constraint weightings are not freely determined but extracted from our experimental data. This removes an entire parameter of variation from the linguist’s control and makes it reflect an objective phenomenon, instead of being a degree of unconstrainedness. Second, these results offer more predictive capacity than might appear at first sight. It is true that the violation costs in numerical form cannot be immediately transferred from one study to another and therefore have only weak predictive capacity, but this is largely a result of their great strength, their exactness. The method is so sensitive to individual characteristics of the experimental materials that changes in linguistic items are clearly reflected in the results. Standard binary grammaticality judgements do not suffer to the same extent because they are so very approximate. One partial solution to this is to blunt this exactness: Keller (2000) does this by making just the violation cost ranking transferable - in our own study this would mean that we should predict that in any study providing an estimate of violation cost, the constraints we tested should rank STR-F << LIM << DAT<ACC << BC-B. Another partial solution is to include a standard reference set of sentences in each experiment, so that all sets of experimental results have a common external reference point with which they can be compared. If far more studies of this type were carried out, rough standard values would no doubt develop, as would consensuses when certain violation types may be expected to occur. Notice that it is would not be possible to compare standard values for “grammatical” and “ungrammatical” judgements for the simple reason that no such values exist. Experience has shown that, released from the constraints of a binary scale, subjects judge sentences such as Mary want to go to school, John hopes the winner and Kim wants any more to eat, which are ungrammatical for different reasons, differently. To summarize, while the precise numerical values of the observed violation costs have only weak predictive capacity, the proportions between violation costs as well as the distribution of violations have strong predictive capacity, since they are readily and robustly replicable, and their reliability is confirmed by the statistical analysis.

Let us also note that Decathlon Model is considerably more constrained than standard OT. The Decathlon Model must produce as output not just the single successful candidate as in OT, but a rough ranking of all sentence types included. This too is to some extent empirically verifiable, since the Decathlon Model predicts that not only the best candidate, but also other close competitors may appear in the linguistic output. This model is therefore readily falsifiable by reference to a syntactically parsed corpus. We would therefore make the claim that account we present here satisfies the criterion of empirical adequacy, that is, it is based on a sound and demonstrably reliable foundation of data and accounts for this data using only constraints motivated elsewhere in the grammar. We think that results such as this should be seen as a confirmation of the soundness of this data-driven approach to syntactic description and theory building.

References

Behaghel O. 1909 Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen 25, 110-142
Everaert M. 1986 The Syntax of Reflexivization. Dordrecht: Foris
Fischer S. 1999 On the integration of cumulative effects into Optimality Theory. ms, University of Tübingen
Frey W. 1993 Syntaktische Bedingungen für die Interpretation. Berlin: AkademieVerlag
Grewendorf G. 1984 Reflexivierungssregeln im Deutschen. Deutsche Sprache 1, 14-30