Dialogue & Discourse 11(1) 1-39 doi: 10.5087/dad.2020.101

The timing of prominence information during the resolution of German personal and demonstrative pronouns

Clare Patterson CLARE.PATTERSON@UNI-KOELN.DE
University of Cologne
Department of German Language and Literature I, Linguistics

Petra B. Schumacher PETRA.SCHUMACHER@UNI-KOELN.DE
University of Cologne
Department of German Language and Literature I, Linguistics

Editor: Massimo Poesio

Submitted 12/2018; Accepted 02/2020; Published online 03/2020

Abstract

German personal and demonstrative pronouns have distinct preferences in their interpretation; personal pronouns are more flexible in their interpretation but tend to resolve to a prominent antecedent, while demonstratives have a strong preference for a non-prominent antecedent. However, less is known about how prominence information is used during the process of resolution, particularly in the light of two-stage processing models which assume that reference will normally be to the most accessible candidate. We conducted three experiments investigating how prominence information is used during the resolution of gender-disambiguated personal and demonstrative pronouns in German. While the demonstrative pronoun required additional processing compared to the personal pronoun, prominence information did not affect resolution in shallow conditions. It did, however, affect resolution under deep processing conditions. We conclude that prominence information is not ruled out by the presence of stronger resolution cues such as gender. However, the deployment of prominence information in the evaluation of candidate antecedents is under strategic control.

Keywords: prominence, anaphora, pronoun resolution, demonstrative, shallow processing, eye-tracking

1 Introduction

Personal pronouns are referentially ambiguous, but they tend to refer to antecedents that are prominent in the prior discourse. This generalisation has led to a great deal of interest in how an entity in discourse becomes more or less prominent. Many different linguistic and non-linguistic factors have been found to contribute to prominence, including (but not limited to) grammatical role (Crawley & Stevenson, 1990; Crawley et al., 1990; Gordon et al., 1993; Järvikivi et al., 2005; Kaiser & Trueswell, 2008), thematic role (Schumacher et al., 2016; Stevenson et al., 1994) and its relation to implicit causality (Garvey, Caramazza & Yates, 1974), word order (Clark & Sengul, 1979; Gernsbacher & Hargreaves, 1988; Järvikivi et al., 2005), and information structure (Almor, 1999; Colonna et al., 2012; Kaiser & Trueswell, 2008). Many languages have a rich set of pronominal devices which are sensitive to different factors. For example, Spanish, Italian, Turkish, Japanese and Greek have null pronouns which tend to refer to prominent entities, but also have one or more overt pronouns that refer to less prominent antecedents.[1] German, like Dutch and Norwegian, does not have null pronouns but does have various demonstratives that can function like personal pronouns by referring to animate entities. While personal pronouns in German tend to refer to prominent antecedents and

[1] Although the division of labour between null and overt pronouns appears to differ between languages (Filiaci et al., 2014).

demonstratives to less prominent ones, the precise division of labour between German personal and demonstrative pronouns is under-researched and still under debate. In this paper we focus on the German personal pronoun "er" and one demonstrative pronoun, "der".

It has long been acknowledged that different anaphoric lexical forms signal different antecedent preferences. For example, Gundel et al. (1993) developed the givenness hierarchy, which relates the use of anaphoric forms to a particular function based on the cognitive status of the antecedent. In this hierarchy, the use of demonstratives signals antecedents that are activated but not in focus, whereas unstressed personal pronouns or zero pronouns refer to something that is in focus (related to the notion of discourse topic). Furthermore, work on English demonstratives has shown that demonstrative pronouns such as *this* differ from *it* not only in the cognitive status or salience of the referent but also the type (and also complexity) of the referent (Brown-Schmidt et al., 2005; Çokal et al., 2018), and links to the idea of a form-specific account (Kaiser & Trueswell 2008). This approach specifies that different referential forms not only seek different types of antecedent but are also differently sensitive to factors affecting the salience or cognitive status of an antecedent, and that accounting for processing of different forms is not possible with a single scale. However, in the current study, we restrict ourselves to two pronominal forms that are both used as independent pronouns referring to animate NP entities. The challenge lies accounting for the resolution preferences and division of labour of these particular forms, which is assumed to be related to prominence.

1.1 Personal and demonstrative pronouns in German

Except for certain instances in colloquial speech, German requires overt pronouns. The least marked form is the unstressed personal pronoun, which is inflected for gender and number as well as case (e.g., nominative singular: *er*/*sie*/*es*; nominative plural: *sie*). More marked variants include the stressed personal pronoun and three different types of demonstrative pronouns (*der*/*die*/*das*, *dieser*/*diese*/*dieses*, *jener*/*jene*/*jenes*). The latter convey greater distance between the speaker and the referent (however this contrast is diminishing; Himmelmann, 1997). The availability of the two proximal forms has been associated with differences in pragmatic function (such as contrast and topic marking) among others (Ahrenholz, 2007). Crucially, demonstrative pronouns in German – unlike English *this* – productively refer to animate entities (see example [1] below).[2] Furthermore, the demonstrative pronoun under investigation here refers to entity-denoting referents and not to propositional content, contrary to *this* and *that* in English (see Çokal et al., 2018; see also Brown-Schmidt et al., 2005 for reference to composite entities). The German demonstrative pronoun for propositional content has a different morphology (*dies*, *das*). In the following, we concentrate on the unstressed personal pronoun *er* and the demonstrative pronoun *der* in German.[3]

To illustrate the different resolution preferences of the two pronoun types, consider the mini-discourse shown in (1) (adapted from Bosch & Umbach, 2007). The second sentence could contain a personal pronoun, shown in bold in (1a), or a demonstrative pronoun, shown in bold in (1b):

(1) Serena wollte mit Venus Tennis spielen.
Serena wanted to play tennis with Venus.

(1a) Doch **sie** war krank.
But she was sick.

(1b) Doch **die** war krank.
But she was sick.

[2] In English it is possible to use a demonstrative such as "this" to refer to human or animate entities in very limited contexts, when they serve as subjects of the verb "to be" in specification contexts such as "This is the woman who saved my life", but in general they can only refer to inanimates when used independently (Stirling & Huddleston, 2002). Using a demonstrative *this/that* instead of *she* in example 1, for instance, would not be possible in English.

[3] Note that the use of *der* as a demonstrative pronoun is fully grammatical in German; it is very frequent in spoken German (with some regional variation) and less frequent, though still attested, in written German (for corpus results see Portele & Bader, 2016).

While the pronouns in (1a) and (1b) are both technically ambiguous, there is a tendency for the personal pronoun in (1a) to refer to the more prominent referent *Serena* and for the demonstrative pronoun in (1b) to refer to the less prominent referent *Venus*. Furthermore, the demonstrative pronoun yields more robust preferences, being more strongly associated with a non-prominent entity, while the personal pronoun tends to be more flexible in its interpretation (Bosch, Katz, & Umbach, 2007; Schumacher et al., 2016).

In particular, it has been a challenge to identify which aspects of prominence are relevant when it comes to the resolution of personal and demonstrative pronouns in German. One proposal is that grammatical role is important: Bouma and Hopp (2006, 2007) claim that subjecthood is an important factor in determining prominence for the personal pronoun *er*, and further that subjecthood is a more important factor than order of mention. When it comes to the comparison between personal and demonstrative pronouns, Bosch et al. (2003) and Bosch et al. (2007) claim that personal pronouns refer to subjects, while the demonstrative *der* refers to non-subjects. However, this position was revised (Bosch & Umbach, 2007) to say that *der* avoids reference to antecedents that are topics. Kaiser (2011a) found that personal pronouns are more flexible than demonstratives, which had a strong object bias in her sentence completion experiment. The experiment also showed that coherence relations did not modulate the interpretation of demonstratives to the same extent as personal pronouns. In Kaiser's study, the contribution of grammatical role, topichood and thematic role was not explored. But the notion that topichood is crucial to understanding the difference between personal and demonstrative pronouns has been taken up and refined by Hinterwimmer (2015), and is confirmed by experimental data from Wilson (2009), who claims that demonstratives are sensitive to topichood (while personal pronouns are sensitive to both topichood and grammatical role). Furthermore, Bosch and Hinterwimmer (2016) incorporate the notion of topichood into the semantic representation of the two pronoun types, with *er* and *der* sharing the same representation except for an "avoid topic" feature for *der*. Effectively, this analysis makes a requirement on the demonstrative to seek out a non-prominent antecedent. This leaves the personal pronoun underspecified and accounts for its greater flexibility in referential preferences while accounting for the difference between the two pronouns.

Evidence from Schumacher and colleagues (Schumacher et al., 2015; 2016; 2017) shows that the findings from Garvey, Caramazza and Yates (1974) and Stevenson et al. (1994) regarding the importance of thematic role for pronoun resolution also applies to German pronouns. Further, they were able to isolate the contribution of thematic role, grammatical role, and word order, and claim that thematic role in some circumstances takes priority over grammatical role in the resolution of German personal and demonstrative pronouns, in making antecedents with the (proto-) agent role more prominent than antecedents with the (proto-) patient role.[4] While an initial study in German by Wilson (2009) using actives and passives to manipulate thematic role showed a null result, as Wilson (2009) acknowledges, this may have been due to the grammatical role hierarchy and thematic role hierarchy pointing in opposite directions. Schumacher and colleagues were able to isolate the contribution of thematic role in German by manipulating verb type, contrasting active accusative and dative experiencer verbs. With active accusative verbs, the nominative-marked argument has the thematic role of (proto-) agent, and the accusative-marked argument has the role of (proto-) patient. With dative experiencer verbs, the dative-marked argument has the thematic role of (proto-) agent and the nominative-marked argument has the role of (proto-) patient. This verb type contrast enabled the hierarchy of grammatical role (nominative subject = most prominent) to be separated from the hierarchy of thematic role (agent = most prominent), because in the dative experiencer verbs, the two hierarchies are not aligned. In Schumacher et al. (2016), an antecedent-selection experiment and two sentence completion experiments, the personal pronoun tended to be resolved to arguments with the thematic role of agent and the demonstrative to arguments with the patient role, irrespective of the grammatical role hierarchy. (Note that manipulations of word order showed that grammatical role was also a relevant parameter, but it appears to be less important than thematic role.) Importantly, in Schumacher's approach, the relevance of thematic role for pronoun resolution does not completely rule out the contribution of other factors such as grammatical role. Indeed, grammatical role may be an important factor when the grammatical role of the pronoun is also considered, where grammatical role parallelism may exert an influence (Sauermann & Gagarina, 2017). Rather, it is proposed that these factors form prominence hierarchies which interact to determine the appropriate referent. Personal and

[4] See Dowty (1991) for a discussion of proto-agent and proto-patient roles.

demonstrative pronouns appear to respond differently to the prominence hierarchies, with personal pronouns referring to the most prominent antecedent and demonstrative pronouns avoiding reference to the most prominent antecedent, referring instead to an antecedent with a lower prominence ranking.

The differing preferences between personal and demonstrative pronouns are also evident during online processing (see Schumacher et al., 2017 for visual world eye tracking data). Furthermore, an event-related brain potential study (Schumacher et al., 2015) revealed two different neurocognitive profiles for personal and demonstrative pronouns in German. The latter evoked a more pronounced negative deflection followed by an enhanced positivity. These two effects have been associated with differing discourse functions. The negative deflection for the demonstrative (compared to the personal pronoun) was associated with the cost of linking of the pronoun with an antecedent. In the case of the demonstrative *der*, this follows the requirement for a less prominent antecedent, with the cost arising either from the process of excluding a prominent antecedent, or from the retrieval of a less prominent entity (see also Burkhardt, 2006; Nieuwland & Van Berkum, 2006; Streb, Hennighausen, & Rösler, 2004 for enhanced negativity effects when anaphor–antecedent linking is more demanding; see Çokal et al., 2018 for increased processing demands associated with propositional antecedents compared to NP antecedents). The enhanced positivity effect, on the other hand, was associated with forward-looking discourse processes. The demonstrative pronoun signals the potential for a thematic shift, unlike the personal pronoun which serves to maintain the existing topic. The shift in attention for the demonstrative elicits an enhanced positivity, as previously observed for topic shift and contrastive focus (Hirotani & Schumacher, 2011; Hung & Schumacher, 2014; Wang & Schumacher, 2013).

Until now, studies contrasting the online processing of personal and demonstrative pronouns in German have used paradigms where the pronouns are ambiguous. When no other cues are available, the calculation of prominence hierarchies may be particularly important in the pronoun resolution process, because they may be the only factor guiding antecedent choice (particularly in the case of the demonstrative, with the requirement to seek out a non-prominent antecedent). However, the role of prominence information is less clear when pronouns are unambiguous, that is, when only a single entity from the prior discourse is a suitable candidate antecedent. In these situations, prominence information, rather than guiding the choice of antecedent, may be used instead to evaluate the felicity of the antecedent that has already been selected. Does prominence information still play an important role in these situations, and is the information used immediately? The current paper examines how prominence hierarchies are used when German personal and demonstrative pronouns are disambiguated by gender cues.

1.2 Gender, processing depth and two-stage pronoun resolution

Many pronoun resolution studies have made use of gender as a disambiguating cue (Badecker & Straub, 2002; Felser & Cunnings, 2012; Sturt, 2003, among many others). However, the precise way in which gender cues interact with accessibility cues has been debated. (In the current paper we equate the notion of prominence with accessibility: we assume that the most prominent discourse entity is cognitively the most accessible (for discussions about accessibility see Ariel, 1990; Arnold, 1998)). There have been claims that gender is used at an early stage of pronoun resolution, followed by accessibility at a later stage (Crawley et al., 1990; Ehrlich, 1980). It has also been claimed, however, that gender information is not always used immediately (Mcdonald & Macwhinney, 1995) or is used strategically (Garnham, Oakhill, & Cruttenden, 1992). Arnold et al. (2000) investigated gender and accessibility cues to pronoun resolution in two visual world eye tracking experiments, manipulating accessibility via order of mention and pronominalisation. They found that both gender and accessibility contribute immediately to the resolution of the pronoun. These conflicting claims have been somewhat clarified by more recent proposals for two-stage pronominal resolution (Rigalleau, Caplan, & Baudiffier, 2004; Stewart, Holler, & Kidd, 2007), as outlined below.

Away from the debate about prominence/accessibility and pronoun resolution, a different line of research has been concerned with whether this process is always completed. For example, Greene, McKoon and Ratcliffe (1992), in a series of probe recognition tasks, demonstrated that a unique referent for a pronoun is not always identified during processing. This finding accords with the shallow processing hypotheses according to which linguistic input is not always fully processed leading to underspecified and sometimes even incorrect interpretations (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007; Sanford & Sturt, 2002). With respect to pronouns, several studies have

shown that whether full interpretation of a pronoun takes place depends on the level of engagement that participants have in the task; for example, Love and McKoon (2011) demonstrated that simply increasing the length of a story text from four to eleven lines increased participants' engagement to the extent that pronoun resolution was successfully completed. A shallow level of processing may be induced in an experimental setting where participants are required to read a number of short texts which are unrelated to each other.

Incremental models of pronoun resolution which assume two processing stages can incorporate the possibility that pronouns may be left underspecified during processing. This has been explicitly incorporated in Rigalleau et al.'s (2004) two-stage pronoun resolution model. This model also accounts for the conflicting findings discussed above for the timing of gender and accessibility information. In the model, the first stage of pronoun resolution consists of co-indexation. Gender cues from the pronoun are automatically checked against the most accessible potential antecedent(s) in the prior discourse. The indexing at this stage does not represent a full commitment to a particular antecedent, rather, it is an automatic process that checks whether resolution is possible by identifying potential antecedents. The second stage involves disengagement: the activation of competing antecedents is reduced so that a single referent for the pronoun is identified. Importantly, the disengagement stage is under strategic control. Strategic control was described in Garnham et al. (1992) as a process that participants could "turn on and off, depending on how they perceive their task", and by Rigalleau et al. (2004) as "requiring time and attention". This second disengagement stage, then, is unlikely to take place under shallow processing. The model therefore assumes that under deep processing conditions a unique antecedent can be identified, whereas under shallow conditions the resolution is left underspecified, with only potential matching antecedents identified. This two-stage model is similar to the bonding and resolution model (Garrod & Sanford, 1990; Garrod & Terras, 2000; Sanford, Garrod, Lucas, & Henderson, 1983), in which candidate antecedents are identified in the bonding stage, and then semantic fit, plausibility and contextual appropriateness are checked in a second stage. Both these models have in common that the first stage involves co-indexation or bonding to the most accessible candidate(s)[5,6]. This implicitly assumes that the relative accessibility of potential antecedents has already been computed. For instance, if grammatical role contributes to antecedent accessibility, then the subject of a sentence would be more accessible than an object, leading to higher activation values for the subject (irrespective of whether it is later pronominalized or not). When a pronoun is subsequently encountered, the subject, being highly accessible, would be checked for gender match with the pronoun in the first stage of resolution. While this assumption fits with pronouns that are usually resolved to the most prominent antecedent, such as personal pronouns in English or German, it is less clear how this works when the pronoun is usually resolved to a less prominent antecedent, as in German demonstrative pronouns.

Stewart et al. (2007) extend Rigalleau et al.'s (2004) two-stage model to ambiguous pronouns. In their study, depth of processing was manipulated by means of comprehension questions. They found that under deep processing (where participants are engaged with the task), ambiguous pronouns take longer to process than unambiguous ones. They attribute this ambiguity cost to the longer disengagement process, where participants have to strategically engage in order to deselect one of the potential pronouns. Under shallow conditions, there is no cost of ambiguity. Evidence from their Experiment 2 points to ambiguous pronoun resolution in shallow conditions being delayed until disambiguating information becomes available, and no initial co-indexation stage. However, even under shallow conditions, the pronouns were nonetheless eventually resolved.

The role of prominence hierarchies, as they are used in the resolution of German personal and demonstrative pronouns, has so far not been addressed in relation to two-stage processing models. If we equate prominence information to the notion of accessibility, then prominence hierarchies should contribute to Stage 1, where co-indexation with the most accessible antecedent(s) takes place. But since we know that demonstrative pronouns have a requirement to avoid the most accessible candidate (while personal pronouns are flexible, or tend to be resolved to the most accessible candidate), the felicity of linking German pronouns to more or less accessible (prominent) candidates must be

[5] Garrod and Sanford (1990) call this the antecedent that is in the reader's attentional focus.

[6] In this paper we concentrate on Rigalleau et al.'s (2004) model and its extension by Stewart et al. (2007) since they explicitly incorporate the notion of deep and shallow processing.

evaluated at some point during processing. This assessment process seems to fit better into the description of Stage 2.

1.3 Current study

The current study investigates the processing of the German personal and demonstrative pronouns with respect to the relative prominence of potential antecedents and looks at how this process unfolds when the pronouns are disambiguated based on gender cues.

The purpose of Experiment 1 was to test the processing of the two pronoun types in German in environments where the referent was unambiguous, with the goal of gaining insight into the timecourse over which the two pronouns are processed and the timecourse of implementing the prominence hierarchies of grammatical role and thematic role. The underlying assumption was that the pronouns would be fully interpreted. Foreshadowing the findings of Experiment 1, we found that the pronouns did not appear to be fully interpreted, and we therefore carried out Experiments 2 and 3 to further investigate the interaction of gender cues and the prominence hierarchies under deep processing conditions.

2 Experiment 1

Experiment 1 made use of the verb type distinction between active accusatives and dative experiencer verbs, as per previous studies by Schumacher and colleagues (Schumacher et al., 2015, 2016, 2017) (*see* Introduction). Using these two verb types allows the prominence hierarchy for grammatical role to be distinguished from that of thematic role. The purpose of Experiment 1 was, firstly, to observe the timecourse over which the two prominence hierarchies interact, and secondly, to compare the processing of personal and demonstrative pronouns. In active accusative verbs the grammatical role and thematic role hierarchies are aligned, with both hierarchies pointing to the nominative marked agent argument as the most prominent antecedent. For example, in (2), the first NP (henceforth NP1) *der Reporter*/*die Reporterin* ("the reporter") is nominative marked (highest grammatical role) and is the proto-agent (highest thematic role), while the second NP (henceforth NP2) *die Sprecherin/den Sprecher* "the spokesperson" is accusative marked (lower-ranked grammatical role) and is the proto-patient (lower thematic role). In the dative experiencer verbs, conversely, the hierarchies are not aligned. As can be seen in (3), the NP1 *dem Trainer/der Trainerin* "the trainer" is marked for dative case (lower grammatical role) and is the proto-agent (highest thematic role), while the NP2 *die Lehrerin/der Lehrer* "the teacher" is nominative marked (highest grammatical role) and is the proto-patient (lower thematic role). If the thematic role hierarchy were the *only* relevant parameter in determining prominence of antecedents, there should not be any difference between the two verb types when the pronouns are processed. If, however, both prominence hierarchies influence pronoun resolution (as suggested in Schumacher et al., 2015, 2016, 2017) we expect to see a processing cost when the two prominence hierarchies are not aligned, as in the dative experiencer verbs. This may be demonstrated in later resolution of pronouns for dative experiencer verbs compared to active accusative verbs. Furthermore, in Experiment 1 we wished to test whether resolution of the demonstrative would be prolonged compared to the personal pronoun. This is based on findings that the neurocognitive profiles of the two pronouns differ (Schumacher et al., 2015). It is assumed that the demonstrative, because of the requirement to pick out a less prominent referent and because of the change in expectations about the upcoming discourse, makes higher processing demands than the personal pronoun. If so, the timecourse of resolution for the demonstrative pronoun may be prolonged.

In order to test these assumptions, we used a gender-match paradigm, which allowed us to detect the resolution of the pronouns. The two arguments of the verb (NP1 and NP2), which served as potential referents for the pronoun, were of different grammatical gender. The pronoun matched in gender with only one of the two potential antecedents, allowing us to manipulate resolution to either antecedent. Here, we predict that the preferences for the personal and the demonstrative pronoun should differ. Following previous findings, personal pronouns should be preferentially resolved to antecedents with a proto-agent role (always NP1), or should not have a strong preference, given that some studies have shown that the personal pronoun is more flexible in its interpretation. The demonstratives, conversely, should prefer reference to the antecedent with a proto-patient role (always NP2), given the lexical-semantic requirement to resolve the demonstrative to a non-prominent antecedent. This means that processing should be easier when the personal pronoun refers to NP1 compared to NP2, and for the demonstrative processing should be easier when referring to the NP2

compared to NP1. This should give rise to an interaction between pronoun (personal/demonstrative) and antecedent (NP1/NP2). This interaction should be delayed for the dative experiencer verbs compared to the active accusative verbs if the non-aligned prominence hierarchies lead to additional processing.

These predictions are based on the assumption that both gender information (information about gender match between the pronoun and the referent) and prominence information (resolution preferences based on prominence hierarchies) would contribute to the processing of the pronouns.

2.1 Materials

Experimental items were short German texts comprising two sentences each. The first sentence contained two animate referents (NP1 and NP2) in the main clause. The second sentence contained a masculine pronoun (*er* or *der*, "he").[7] 32 items contained an active accusative verb in the main clause of sentence 1, and 32 items contained a dative experiencer verb in the main clause of sentence 1, giving a total of 64 experimental items across the two verb types.[8] Within each verb type two factors were systematically manipulated: pronoun (*er* versus *der*); and antecedent (NP1 versus NP2). The factor *antecedent* was manipulated via gender match: the gender of the pronoun matched either the gender of the NP1 or the NP2 (the subordinate clause did not contain any masculine nouns). In the active accusative sentences NP1 was always in nominative case, and in the dative experiencer sentences the NP1 was always in dative case (i.e. the base argument order). An example of the conditions is given in (2) and (3) below (note that the idiomatic translation given at the end of each item set covers all four conditions).

In addition, the following aspects of the materials were controlled in order to minimise the variation in eye-movements between and across conditions: there were always four words following the pronoun; the word preceding the pronoun was always *aber* or *doch* ("but", "yet/still"). Nouns in NP1 position and NP2 position did not differ (overall) in frequency, and were matched per item in letter length and syllable length.

(2) **Active-accusative verbs**

2a. *ER, NP1*

Der Reporter wollte die Sprecherin befragen, weil die Konferenz ausfiel. Aber er hatte dann keine Zeit.

Der	Reporter	woll-te	die	Sprecher-in	befragen,
the.M.SG.NOM	*reporter.M*	*want.PST-3SG*	*the.F.SG.ACC*	*spokesperson-F*	*interview.INF*
weil	die	Konferenz	aus-fiel.		
because	*the.F.SG.NOM*	*conference.F*	*out-fall.PST*		
Aber	er	hat-te	dann	kein-e	Zeit.
but	*he*	*have.PST-3SG*	*then*	*no-F*	*time.F*

2b. *ER, NP2*

Die Reporterin wollte den Sprecher befragen, weil die Konferenz ausfiel. Aber er hatte dann keine Zeit.

Die	Reporter-in	woll-te	den	Sprecher	befragen,
the.F.SG.NOM	*reporter.F*	*want.PST-3SG*	*the.M.SG.ACC*	*spokesperson.M*	*interview.INF*
weil	die	Konferenz	aus-fiel.		
because	*the.F.SG.NOM*	*conference.F*	*out-fall.PST*		
Aber	er	hat-te	dann	kein-e	Zeit.
but	*he*	*have.PST-3SG*	*then*	*no-F*	*time.F*

[7] Only masculine singular forms were used; the feminine pronouns are ambiguous for case and number.

[8] Note that there are only a limited number of dative-experiencer verbs in German. In the current experiment six verbs were used and repeated with different contexts to obtain 32 items: *auffallen* ("to catch so. eye"), *entgehen* ("to escape/evade so."), *missfallen* ("to displease so."), *imponieren* ("to impress so."), *behagen* ("to please so."), gefallen ("to please so.").

2c. *DER, NP2*
Die Reporterin wollte den Sprecher befragen, weil die Konferenz ausfiel. Aber der hatte dann keine Zeit.

Die	Reporter-in	woll-te	den	Sprecher	befragen,
the.F.SG.NOM	*reporter.F*	*want.PST-3SG*	*the.M.SG.ACC*	*spokesperson.M*	*interview.INF*
weil	die	Konferenz	aus-fiel.		
because	*the.F.SG.NOM*	*conference.F*	*out-fall.PST*		
Aber	der	hat-te	dann	kein-e	Zeit.
but	*he.DEM*	*have.PST-3SG*	*then*	*no-F*	*time.F*

2d. *DER, NP1*
Der Reporter wollte die Sprecherin befragen, weil die Konferenz ausfiel. Aber der hatte dann keine Zeit.

Der	Reporter	woll-te	die	Sprecher-in	befragen,
the.M.SG.NOM	*reporter.M*	*want.PST-3SG*	*the.F.SG.ACC*	*spokesperson-F*	*interview.INF*
weil	die	Konferenz	aus-fiel.		
because	*the.F.SG.NOM*	*conference.F*	*out-fall.PST*		
Aber	der	hat-te	dann	kein-e	Zeit.
but	*he.DEM*	*have.PST-3SG*	*then*	*no-F*	*time.F*

"The reporter wanted to interview the spokesperson, because the conference was cancelled. But he didn't have time then."

(3) **Dative-experiencer verbs**

3a. *ER, NP1*
Dem Trainer hatte die Lehrerin imponiert, und zwar seit dem letzten Sportfest. Doch er wollte das nicht zugeben.

Dem	Trainer	hat-te	die	Lehrer-in	imponiert,
the.M.SG.DAT	*trainer.M*	*have.PST-3SG*	*the.F.SG.NOM*	*teacher-F*	*impress.PTCP*
und	zwar	seit	dem	letzt-en	Sportfest.
and	*indeed*	*since*	*the.N.SG.DAT*	*last-N.SG.DAT*	*sports.day.N*
Doch	er	woll-te	das	nicht	zugeben.
but	*he*	*want.PST-3SG*	*that*	*not*	*admit.INF*

3b. *ER, NP2*
Der Trainerin hatte der Lehrer imponiert, und zwar seit dem letzten Sportfest. Doch er wollte das nicht zugeben.

Der	Trainer-in	hat-te	der	Lehrer	imponiert,
the.F.SG.DAT	*trainer-F*	*have.PST-3SG*	*the.M.SG.NOM*	*teacher.M*	*impress.PTCP*
und	zwar	seit	dem	letzt-en	Sportfest.
and	*indeed*	*since*	*the.N.SG.DAT*	*last-N.SG.DAT*	*sports.day.N*
Doch	er	woll-te	das	nicht	zugeben.
but	*he*	*want.PST-3SG*	*that*	*not*	*admit.INF*

3c. *DER, NP2*
Der Trainerin hatte der Lehrer imponiert, und zwar seit dem letzten Sportfest.
Doch der wollte das nicht zugeben.

Der	Trainer-in	hat-te	der	Lehrer	imponiert,
the.F.SG.DAT	*trainer-F*	*have.PST-3SG*	*the.M.SG.NOM*	*teacher.M*	*impress.PTCP*
und	zwar	seit	dem	letzt-en	Sportfest.
and	*indeed*	*since*	*the.N.SG.DAT*	*last-N.SG.DAT*	*sports.day.N*
Doch	der	woll-te	das	nicht	zugeben.
but	*he.DEM*	*want.PST-3SG*	*that*	*not*	*admit.INF*

3d. *DER, NP1*
Dem Trainer hatte die Lehrerin imponiert, und zwar seit dem letzten Sportfest.
Doch der wollte das nicht zugeben.

Dem	Trainer	hat-te	die	Lehrer-in	imponiert,
the.M.SG.DAT	*trainer.M*	*have.PST-3SG*	*the.F.SG.NOM*	*teacher-F*	*impress.PTCP*
und	zwar	seit	dem	letzt-en	Sportfest.
and	*indeed*	*since*	*the.N.SG.DAT*	*last-N.SG.DAT*	*sports.day.N*
Doch	der	woll-te	das	nicht	zugeben.
but	*he.DEM*	*want.PST-3SG*	*that*	*not*	*admit.INF*

"The teacher had impressed the trainer, in particular since the last sports day. But he didn't want to admit it."

The 64 experimental items were interspersed with 96 fillers (64 containing feminine pronouns) and divided over 8 lists in a Latin-square design.

2.2 Participants

Data was collected from 35 native German speakers (10 male), age range 19-40 years, of whom 32 were included in the analysis. (Two participants were excluded for excessive track loss, and one participant was excluded due to low accuracy (<70%) on the comprehension questions.) No participant reported any language disorders. All participants gave their consent and received a small fee or course credit for participation.

2.3 Procedure

Participants were seated with their eyes 70cm from the computer screen displaying the text, with their head supported by a chin-rest and forehead-rest. Texts were displayed on the screen in a black font (20pt) on a white background using the Courier New font. Texts were displayed over two lines. The line break was in the subordinate clause of the first sentence so that the second sentence always started in the middle of the second line, away from the line break. Participants were asked to read sentences silently from computer screen at their normal reading rate and to answer comprehension questions on gamepad, while their eye movements were recorded using the Eyelink 1000 (SR Research) desktop mount. Comprehension questions (Y/N) followed half of experimental items and approx. quarter of the fillers (58 questions in total). Six questions directly probed the referent of the pronoun, and five no-questions probed aspects of the pronoun sentence. The experiment took around 45 minutes, and participants were paid a small amount or given course credit for their participation.

2.4 Analysis

2.4.1 Regions of Interest

For the analysis, the sentences were divided into regions. The pronoun region consisted of the first two words of the second sentence: *aber er* or *aber der* ("but he"). The spillover region consisted of the two words following the pronoun.

2.4.2 Data cleaning procedure

If an individual fixation was shorter than 80ms, it was merged with a nearby fixation (within 1 degree visual angle). If there was no nearby fixation to merge it with, the fixation was deleted. Individual fixations longer than 1200ms were deleted. In a given trial, any regions that were skipped during first-pass reading were removed from the analysis and counted as missing data. Skipping rates were 6.15% in the pronoun region and 2.83% in the spillover region.

2.4.3 Data analysis procedure

In the two regions the following reading measures were calculated: first-pass times (summed duration of fixations in a region before exiting it for the first time); right-bound reading times (summed duration of fixations in a region before exit to a later region); rereading times (summed duration of fixations minus the first-pass times; when no rereading took place, this was counted as missing data, making the rereading times a contingent measure); and total viewing times (summed duration of all fixations in a region). Linear mixed-effects models to analyse the data per verb type/region/measure using the package lmerTest version 2.0-33 (Kuznetsova, Brockhoff, & Christensen, 2016) in the R statistical program (R Core Team, 2017). The data was transformed before submitting it to the model. The transformation was determined with the Box-Cox procedure (Box & Cox, 1964) using recommendations from Osborne (2010). This resulted in a reciprocal square-root transformation for the pronoun region and a log transformation in the spillover region.[9] The fixed part of each model contained the sum-coded factors pronoun (*der*; *er*) and antecedent (*NP1; NP2*) and the (centred) trial number, as well as the interactions between pronoun and antecedent. (The trial number was included to control for effects over the course of the experiment; the output was interpreted in an exploratory analysis, see section 2.7.) The random part of the model contained random intercepts for participant and item. The inclusion of by-item and by-participant random slopes for pronoun and antecedent were determined using the function rePCA from the package RePsychLing (Baayen, Bates, Kliegl, & Vasishth, 2015).

2.5 Predictions

Assuming that (proto-) agenthood plays an important role in the selection of an antecedent during pronoun resolution in German (Schumacher et al., 2015; 2016; 2017), in both the active accusative and the dative experiencer items, it should be easier to process *er* when it refers to NP1 (proto-agent), compared to NP2 (proto-patient). Conversely, it should be easier to process *der* when it refers to NP2 compared to NP1. This should lead to an interaction of pronoun and antecedent: ER-NP1 will have shorter reading times than ER-NP2 and DER-NP2 will have shorter reading times than DER-NP1. If the misalignment of prominence hierarchies in the dative experiencer verbs leads to additional processing demands, the interaction of pronoun and antecedent in these items should appear in a later measure (rereading times) or a later region (spillover region) than the interaction for the active accusative items.

Building on the claim that the demonstrative pronoun is more rigid in its interpretive preferences and rejects the most prominent entity, enhanced processing costs are predicted. If the demonstrative pronoun *der* exerts more processing demands than the personal pronoun *er*, we will see overall higher reading times for *der* than *er*: this should be detected in the cumulative measure (total viewing times) and may be visible in either earlier measures (first-pass times; right-bound reading times) or later ones (rereading times). Here, it is also possible that the processing for *der* is prolonged such that the expected NP2-NP1 difference for *der* appears in later measures or a later region than the NP2-NP1 difference for *er*.

[9] Note that this results in the effect directions being reversed for the pronoun region.

2.6 Results

2.6.1 Comprehension questions

One participant scored below 70% in the comprehension questions and was removed from the analysis. For the remaining participants (n=32), overall accuracy was 83% (range 72-90%).

2.6.2 Eyetracking results

The means for each region and measure in the active-accusative items are shown in Table 1. Table 2 shows the outcome of the statistical models (effects of trial were included only as a control measure and are therefore not shown in this table – but see the exploratory analysis below).

	First-pass times	Right-bound reading times	Rereading times	Total-viewing times
PRONOUN REGION				
	Mean	Mean	Mean	Mean
DER, NP1	339 (196)	355 (223)	357 (192)	458 (289)
DER, NP2	332 (186)	342 (203)	299 (156)	406 (251)
ER, NP1	308 (169)	316 (182)	308 (244)	376 (250)
ER, NP2	293 (154)	305 (167)	269 (145)	358 (194)
SPILLOVER REGION				
	Mean	Mean	Mean	Mean
DER, NP1	360 (181)	401 (198)	379 (328)	491 (276)
DER, NP2	371 (181)	405 (207)	360 (211)	485 (271)
ER, NP1	359 (189)	385 (201)	379 (259)	477 (278)
ER, NP2	355 (195)	375 (203)	329 (213)	446 (246)

Table 1. Means (in ms) for the pronoun and spillover regions in the active accusative items in Experiment 1. Standard deviations shown in parentheses.

	First-pass times		Right-bound reading times		Rereading times		Total viewing times	
Effect	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*
PRONOUN REGION								
Pronoun	-1.239e-03 (4.124e-04)	-3.005**	-1.344e-03 (3.794e-04)	-3.542**	-2.072e-03 (1.401e-03)	-1.480	-1.710e-03 (5.080e-04)	-3.366**
Antecedent	-3.901e-04 (3.180e-04)	-1.227	-3.309e-04 (3.232e-04)	-1.024	-1.555e-03 (8.974e-04)	-1.733(*)	-7.343e-04 (4.254e-04)	-1.726(*)
Pronoun x Antecedent	1.841e-04 (3.179e-04)	0.579	2.380e-05 (3.231e-04)	0.074	-8.863e-04 (8.936e-04)	-0.992	-5.329e-04 (3.525e-04)	-1.512
SPILLOVER REGION								
Pronoun	1.634e-02 (2.080e-02)	0.786	3.137e-02 (1.661e-02)	1.888(*)	1.585e-02 (3.190e-02)	0.497	3.187e-02 (1.497e-02)	2.128*
Antecedent	-6.690e-03 (1.517e-02)	-0.441	2.092e-03 (1.558e-02)	0.134	1.481e-02 (3.186e-02)	0.465	1.676e-02 (1.496e-02)	1.120
Pronoun x Antecedent	-1.390e-02 (1.386e-02)	-1.003	-8.734e-03 (1.323e-02)	-0.660	-1.652e-02 (3.149e-02)	-0.525	-2.910e-03 (1.496e-02)	-0.195

Table 2. Model outputs for first-pass times, right-bound reading times, rereading times and total viewing times in the pronoun and spillover regions, active-accusative items, Experiment 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; (*) $p < 0.1$.

For the active accusative items, there is a main effect of pronoun in the pronoun region in first-pass and right-bound reading times, with *der* taking longer than *er*. The same effect is also reflected in the total viewing times for both the pronoun and the spillover regions. Additionally, there are marginal effects of antecedent in rereading and total-viewing times in the spillover region, with longer reading times when the pronouns refer to NP1.

The means for each region and measure in the dative-experiencer items are shown in Table 3. Table 4 shows the outcome of the statistical models.

	First-pass time	Right-bound reading time	Rereading time	Total-viewing time
PRONOUN REGION				
	Mean	Mean	Mean	Mean
DER, NP1	361 (228)	383 (255)	370 (263)	488 (319)
DER, NP2	364 (225)	371 (232)	391 (261)	495 (311)
ER, NP1	305 (177)	319 (195)	349 (278)	395 (296)
ER, NP2	311 (172)	321 (195)	281 (168)	372 (226)
SPILLOVER REGION				
	Mean	Mean	Mean	Mean
DER, NP1	387 (216)	419 (232)	373 (259)	512 (294)
DER, NP2	353 (174)	405 (199)	354 (201)	493 (246)
ER, NP1	365 (190)	386 (192)	350 (211)	474 (251)
ER, NP2	353 (195)	384 (224)	376 (239)	469 (276)

Table 3. Means (in ms) for the pronoun and spillover regions in the dative experiencer items in Experiment 1. Standard deviations shown in parentheses.

	First-pass times		Right-bound reading times		Rereading times		Total viewing times	
Effect	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*
PRONOUN REGION								
Pronoun	-1.857e-03 (3.913e-04)	-4.745***	-1.887e-03 (4.279e-04)	-4.410***	-2.675e-03 (9.579e-04)	-2.793**	-3.026e-03 (3.894e-04)	-7.772***
Antecedent	2.985e-04 (3.398e-04)	0.879	-2.698e-05 (3.332e-04)	-0.081	-9.335e-04 (8.681e-04)	-1.075	5.799e-05 (4.144e-04)	0.140
Pronoun x Antecedent	-2.141e-04 (3.401e-04)	-0.630	-3.014e-04 (3.334e-04)	-0.904	1.586e-03 (8.696e-04)	1.824(*)	-5.772e-06 (3.712e-04)	-0.016
SPILLOVER REGION								
Pronoun	1.391e-02 (1.400e-02)	0.994	3.350e-02 (1.639e-02)	2.044*	-5.552e-03 (3.052e-02)	-0.182	3.277e-02 (1.607e-02)	2.039(*)
Antecedent	2.700e-02 (1.399e-02)	1.930(*)	1.062e-02 (1.333e-02)	0.797	-3.780e-03 (3.059e-02)	-0.124	1.155e-02 (1.458e-02)	0.792
Pronoun x Antecedent	7.419e-03 (1.398e-02)	0.531	-4.347e-03 (1.332e-02)	-0.326	1.381e-02 (3.053e-02)	0.452	-3.578e-03 (1.457e-02)	-0.245

Table 4. Model outputs for first-pass times, right-bound reading times, rereading times and total viewing times in the pronoun and spillover regions, dative experiencer items, Experiment 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; (*) $p < 0.1$.

For the dative experiencer items, there is a main effect of pronoun in all measures of the pronoun region, with *der* taking longer than *er*. The same effect is also reflected in the right-bound reading times in the spillover region. Additionally, there is a marginal main effect of antecedent in the first-pass times of the spillover region with longer reading times for NP1, and in the pronoun region rereading times there is a marginal pronoun by antecedent interaction.

2.7 Exploratory analysis

There is a persistent main effect of pronoun throughout the experiment, which could be associated with differences in resolution processes but also more low-level differences between *er* and *der*. There are several word-level factors which may make the demonstrative harder to process than the personal pronoun. On its own, the word *der* is ambiguous between a definite article and a demonstrative, with the demonstrative pronoun being the less frequent of these uses; it is less frequent (in its pronominal usage) than the personal pronoun, and it tends to appear more in the spoken than in the written modality. We would expect frequency and ambiguity to affect reading times, but if they are the sole basis for the pronoun effect we would expect them to diminish over the course of the experiment. This is because participants tend to adapt very quickly to the local conditions of an experimental setting; participants should get used to *der* appearing as a demonstrative pronoun leading to a lowering of the associated processing cost. This was directly tested in this exploratory analysis (not pre-planned) by checking whether trial number interacted with the effect of pronoun. The same statistical models as in the main analysis were rerun, this time adding the interaction with trial. Outputs for the pronoun by trial interaction are shown in Appendix A. There was only one pronoun by trial interaction (rereading times, spillover region for the active accusatives). Otherwise there were no interactions of the pronoun effect with trial.

2.8 Discussion

Main effects of pronoun, with longer reading times for *der*, are seen throughout the experiment across both active accusative and dative experiencer verbs, most strongly visible in the pronoun region but also reflected in the spillover region. This effect is seen in both early and later eyetracking measures. The expected pronoun by antecedent interaction was not detected, except marginally in the rereading times for the dative experiencer verbs.

Based on findings from Schumacher et al. (2015), we expected to find that the demonstrative pronoun would have a processing cost relative to the personal pronoun. This was seen in the main effects of pronoun throughout the experiment. In order to rule out that these effects are due to low-level differences in frequency or ambiguity of form, we checked whether the pronoun effects diminish over the course of the experiment. Given that the pronoun effect largely did not interact with trial (as shown in the exploratory analysis), we find it unlikely that the pronoun effect comes from low-level differences between the two pronouns. Our prediction of a processing cost for *der* was based on the distinct discourse functions of the two pronouns. One function is the requirement for the demonstrative pronoun to pick out a non-prominent referent as its antecedent; this is associated with higher processing demands, either because a prominent antecedent has to be excluded (in contrast to the unspecific resolution instruction associated with the personal pronoun), or because there is a cost for retrieving a less accessible antecedent. The other function is the change in expectations about the upcoming discourse, because demonstratives are associated with a topic shift. However, both of these functions pertain to the demonstrative referring to the non-prominent antecedent (NP2 in Experiment 1). But in our experiment, reference to NP1 and NP2 was systematically manipulated. Therefore the discourse functions would only explain the processing cost for the demonstrative if there was also an interaction between pronoun and antecedent. An alternative explanation for the cost for the demonstrative which is not associated with its eventual referent is the unexpectedness of a form which signals a referential shift instead of thematic maintenance, regardless of the eventual referent (see Portele & Bader, 2016 for production data in German that show a low likelihood of demonstrative pronoun continuations).

An alternative explanation for the extra cost associated with the *der* is the one letter length difference between the DER region (7 characters) versus the ER region (6 characters). We think that this is unlikely to be the sole reason for the cost associated with *der*, given the persistence and strength of the effect. Length differences are likely to give rise to differences in landing position and gaze

duration (Rayner, 1998) but are not normally also associated with increased processing downstream, especially when the difference is only one letter.

As discussed in the introduction, previous studies investigating the interpretation of personal and demonstrative pronouns have shown that personal pronouns tend to refer to entities with a proto-agent role (or to be flexible in their interpretation) while demonstratives tend to refer to entities with a proto-patient role. We expected that these interpretive preferences would be reflected in the reading record as an interaction between pronoun and antecedent. However, this pattern was not detected. We are therefore unable to assess whether the alignment of grammatical role and thematic role prominence hierarchies, which are aligned in the active accusative verbs but not aligned in the dative experiencer verbs, had an impact on processing times for pronoun resolution. The lack of a pronoun by antecedent interaction is surprising, given how robust the interpretation preferences have been in previous offline studies and in visual-world eye-tracking (Bosch et al., 2007; Schumacher et al., 2016, 2017). By using gender cues to resolve the pronoun to one antecedent or another, participants in the current experiment were presented with sentences that should have been infelicitous (in particular the DER-NP1 condition, but possibly also ER-NP2 requiring the less preferred resolution site), yet this was not evident in the reading times. Even if the personal pronoun is much more flexible than the demonstrative in interpretive preferences (Bosch et al., 2007; Schumacher et al., 2016), which may minimise the difference between the personal pronoun conditions, there should still have been an effect for the demonstrative. We consider two (related) possible reasons why the interaction was not detected.

One possibility is the way in which gender cues interact with cues that arise from the discourse structure (the prominence hierarchies). When presented with an ambiguous pronoun (no gender cues), as in previous studies, the act of resolving the pronoun may make use of any and all available cues to guide resolution. This means that information from the relative prominence of antecedents becomes important because it is able to guide resolution. Conversely, when gender cues are sufficient to identify a single antecedent, as in the current experiment, prominence cues may be less important because they are not determining reference in any way (cf. Crawley et al., 1990; Ehrlich, 1980; Garnham & Oakhill, 1985; Arnold et al., 2000 for personal pronouns). It is possible that gender cues completely overrule the prominence hierarchies because they are sufficient to identify a single referent, therefore no further processing is needed. In this way, even if the personal pronoun refers to the least prominent antecedent or the demonstrative to the most prominent, it will not affect the processing so long as a unique antecedent has been identified through gender cues. Note that this possibility, if correct, would have implications for a two-stage model of pronoun resolution. It implies the following: if, at the indexing stage (Stage 1), a single antecedent had been identified through gender cues, then evaluation of the link at Stage 2, at least with respect to prominence cues, is effectively cancelled.

The detection of a pronoun by gender interaction rests on the assumption that the pronouns are fully processed during the experiment. However, it is possible that participants were engaged in shallow processing. In the current experiment, the participants had to frequently answer comprehension questions, but most of these questions did not probe the referent of the pronoun. Previous research has shown that, under similar conditions, participants did not always fully resolve pronouns (Garnham et al., 1992; Greene et al., 1992; Rigalleau et al., 2004). An initial indexing stage may take place (and make use of gender cues, as discussed in the introduction), but a more strategic evaluation of candidate antecedents may not take place.

Under both scenarios, the use of information from prominence hierarchies is restricted or blocked. But from the current results it is unclear whether this is due to the task demands, or whether strong cues such as gender more generally limit the use of other, more subtle cues to pronoun interpretation. In order to decide between these two possibilities, Experiment 2 was carried out.

3 Experiment 2

Experiment 2 was an untimed rating task, using a subset of materials from Experiment 1. By asking participants to rate each item, we expected them to engage in deeper, more strategic processing than in Experiment 1. This enabled us to test whether the presence of gender cues effectively blocks the use of prominence information, or whether prominence information can affect pronoun resolution if the task demands encourage deeper processing.

3.1 Materials

The materials consisted of 32 experimental items (16 active accusative, 16 dative experiencer) and 28 fillers which were a subset of the experimental and filler items from Experiment 1. Each experimental item appeared in 4 conditions, as per Experiment 1. To ensure sufficient variation in the verbs for the dative experiencer conditions, each verb used in Experiment 1 appeared at least twice, and four verbs (*auffallen* ("to catch so. eye"), *entgehen* ("to evade/escape so."), *missfallen* ("to displease so."), *gefallen* ("to please so.")) appeared three times.

The 28 fillers were a subset of the fillers from Experiment 1, 16 of which contained a feminine pronoun (*sie* or *die*) to balance out the masculine pronouns appearing in the experimental items; the remaining 12 fillers did not contain any pronoun. Half of the fillers were adjusted to make them sound unnatural or implausible. This was to encourage participants to use a wide range in the rating scale.

3.2 Participants

Data was collected from 42 native speakers of German (8 male), age range 18-38 years; one participant was bilingual (German/English). No participant reported any language disorders. All participants gave their consent and received course credit for participation.

3.3 Procedure

The experimental items were mixed with the fillers and distributed over four lists in a Latin-square design. Each participant saw only one list. Access to the experiment was via a link, and participants completed it remotely via the Qualtrics survey platform (Qualtrics, Provo, UT). Participants were instructed to read each text carefully and then to rate how good each text sounds from 1 *sehr schlecht* ("very bad") to 7 *sehr gut* ("very good"). Evaluations were given by clicking on stars under each text.

3.4 Predictions

If the presence of gender cues completely blocks the use of information from the prominence hierarchies, there should be no difference in ratings whether the pronouns refer to NP1 or to NP2. If, however, prominence information is used despite the presence of gender cues, then we expect an interaction of pronoun and antecedent. Specifically, ER-NP1 should be rated significantly higher than ER-NP2*;* and DER-NP2 should be rated significantly higher than DER-NP1.

3.5 Data analysis

Scores from 1-7 for each participant and item combination (including fillers) were converted to z-scores, as is generally recommended for judgment data (Schütze & Sprouse, 2014) to account for participants' variation in the use of the scale.

The filler items were analysed to ensure that the task had been carried out as expected. Mean z-scores for the adjusted (implausible) filler items were lower than the scores for the unadjusted items (-0.40 versus 0.83), and this difference was significant when tested in a linear mixed-effects model (t=7.391, p<.001). In addition, each participant's mean z-score for the adjusted fillers was lower than their mean z-score for the unadjusted items, confirming that they all performed the task as expected.

Z-scores for the experimental items were then analysed using a linear mixed-effects model using the package lmerTest version 2.0.33 (Kuznetsova et al., 2016) in the R statistical program (R Core Team, 2017). The fixed part of the model contained the sum-coded factors pronoun (*er*; *der*), antecedent (*NP1; NP2*), and verb type (*active accusative*; *dative experiencer*), and the (centred) trial number, as well as the interactions between them. The random part of the model contained random intercepts for participant and item. The inclusion of random slopes for pronoun, antecedent and verb type were determined using the function rePCA from the package RePsychLing (Baayen et al., 2015).

3.6 Results

Table 5 shows the mean raw and z-scores per condition and verb type.

	Active accusative verbs		Dative experiencer verbs	
	Raw scores	Z-scores	Raw scores	Z-scores
ER, NP1	4.40 (2.03)	0.19 (0.88)	3.30 (1.88)	-0.35 (0.84)
ER, NP2	4.08 (2.00)	0.05 (0.80)	3.43 (1.80)	-0.33 (0.84)
DER, NP1	3.53 (2.00)	-0.24 (0.88)	2.68 (1.77)	-0.68 (0.72)
DER, NP2	4.58 (1.90)	0.27 (0.88)	3.18 (1.83)	-0.44 (0.76)

Table 5. Mean raw scores and z-scores per condition and verb type for Experiment 2. Standard deviations shown in parentheses.

The model showed main effects of pronoun (t=-3.340, p=.001), antecedent (t=-2.966, p=.005) and verb type (t=5.517, p<.001). There was a significant interaction of pronoun and antecedent (t=-5.427, p<.001) and a significant interaction of pronoun, antecedent and verb type (t=-2.583, p=.010).

Data was split by verb type and pronoun to follow up the interactions (p-values were Bonferroni-corrected to adjust for the four tests). For the active accusative verbs, the personal pronoun *er* showed no effect of antecedent (t=1.633, p=.107), whereas for the demonstrative pronoun *der* there was a significant main effect of antecedent (t=-4.660, p<.001), with significantly lower scores for the NP1 condition versus NP2. For the dative experiencer verbs, the personal pronoun *er* showed no effect of antecedent (t=-0.108, p=.915), whereas for the demonstrative pronoun *der* there was a main effect of antecedent (t=-2.339, p=.025), with lower scores for the NP1 condition versus NP2.

3.7 Discussion

Based on the results of Experiment 1, it was possible that the presence of gender cues that identified a single antecedent blocked the use of information from prominence hierarchies. The purpose of Experiment 2 was to test whether or not this is the case in an untimed experiment in which participants were asked to evaluate the items from Experiment 1.

Experiment 2 showed that participants were sensitive to prominence hierarchies despite the presence of gender cues; while ratings did not differ significantly for the personal pronoun *er*, they did for the demonstrative *der*, such that reference to the more prominent antecedent (NP1) was rated lower than reference to the less prominent antecedent (NP2). This finding is in line with previous studies which have shown stronger preferences for *der* compared to *er*, and that *der* is preferentially resolved to an antecedent that is not the most prominent one.

Scores for sentences containing the demonstrative pronoun were not in general lower than those containing the personal pronoun. This suggests that in general, the demonstrative pronoun does not sound unnatural in these contexts compared to the personal pronoun. We interpret this as further evidence that the longer reading times for the demonstrative in Experiment 1 are due to a specific processing burden, not captured by an offline rating task, rather than a general dispreference for the demonstrative.

Furthermore, the scores for the active accusative sentences were higher overall than those for the dative experiencer sentences. This could indicate that the misalignment of thematic role and grammatical role prominence hierarchies in dative experiencer verbs, as predicted for Experiment 1 but not confirmed, does indeed have an impact on resolution. However, it is not possible in this experiment to rule out alternative explanations such as a dispreference for having a dative argument in first position or the lower overall frequency of these verbs compared to the active accusatives.

Importantly, the findings of Experiment 2 rule out the possibility that gender cues completely block the use of prominence information. We attribute the difference between the findings for Experiments 1 and 2 to the difference in task: in Experiment 1, participants were not required to resolve the pronoun to answer (most of) the comprehension questions, nor were they asked to evaluate the sentence in any way. These are the conditions under which shallow processing may take place. In Experiment 2, the rating task meant that participants had to engage in deeper processing in order to make a meaningful judgment. It therefore seems likely that the lack of prominence effects in Experiment 1 was due to shallow processing, and that prominence hierarchies do indeed play a role when participants are forced to engage more deeply with pronoun resolution. This result further suggests that the use of prominence information for evaluation of reference is under strategic control.

While prominence information may be used during an offline evaluation, the timecourse over which the interaction between gender cues and prominence information plays out remains unclear. Note that the two-stage processing models discussed in the Introduction do not specify whether prominence information is only part of indexing, in that it contributes to the accessibility of certain candidates, or whether the prominence hierarchies also contribute more strategically to the evaluation of candidate antecedents, based on preferences for candidates ranked higher (in the case of personal pronouns) or lower (in the case of demonstrative pronouns) in the prominence hierarchies. We carried out Experiment 3 to investigate the timing of prominence information during pronoun resolution.

4 Experiment 3

The findings of Experiments 1 and 2 seem to indicate that the use of prominence information in German pronoun resolution is at least partly under strategic control of the participants, with its effects showing up during deeper processing but not under shallow processing conditions. The purpose of Experiment 3 was to investigate the timing of prominence information during the pronoun resolution process, by comparing the processing of pronouns that are (temporarily) ambiguous versus unambiguous with respect to gender cues under deep processing conditions. In this experiment, comprehension questions directly probed the referent of the pronoun in half the experimental items, such that deep processing should take place throughout the experiment. We expect that, during deep processing, prominence information will be used to evaluate candidates during Stage 2 of pronoun resolution, resulting in the pronoun by antecedent interactions that we expected, but did not find, in Experiment 1. Furthermore, we expect that the timing of the use of prominence information will differ between (temporarily) ambiguous conditions and unambiguous conditions. In the unambiguous conditions, gender cues point to a single potential antecedent during the indexing stage, so evaluation of the link with respect to prominence (Stage two) can begin straight away. In the ambiguous conditions, gender cues do not narrow down the set of potential antecedents for the pronoun. Disambiguating information becomes available further on in the sentence, so we expect that at this later point, a full stage two evaluation takes place in the ambiguous conditions.[10] This means that pronoun by antecedent interactions should be seen earlier in unambiguous conditions compared to (temporarily) ambiguous conditions where the gender cues do not identify a single candidate antecedent.

4.1 Materials

Experimental items were 64 short German texts comprising two sentences each. The first sentence comprised a main and subordinate clause, with two animate referents (NP1 and NP2) in the main clause and an active-accusative main verb. The second sentence started with a time adverbial (e.g. *später* "later"*, kurz darauf* "shortly afterwards"*, danach* "afterwards") followed by an auxiliary verb and a masculine pronoun (*er* or *der* "he") as the subject of the main verb, followed by a buffer region containing an adverbial phrase. Following the buffer, either the NP1 or the NP2 from the first sentence was repeated as the object; the repeated NP could not therefore be the referent of the pronoun. The buffer containing the adverbial phrase before the repeated NP was always at least 10 characters long, to ensure that the disambiguating information from the repeated NP was not visible when the pronoun was initially encountered.

Three factors were manipulated: pronoun (*er* versus *der*); antecedent (NP1 versus NP2) and disambiguation (early versus late), giving rise to eight conditions. In the *early disambiguation* conditions, the referent of the pronoun was disambiguated through gender match, as per Experiments 1 and 2; the gender of the two NPs always differed (one masculine, one feminine). The position of the gender matching (masculine) antecedent was manipulated (NP1 or NP2) to create the two antecedent conditions. In the *late disambiguation* conditions, the gender of NP1 and NP2 was always the same

[10] We remain ambivalent on precisely what happens at the point of the pronoun in the (temporarily) ambiguous conditions. Following Stewart et al. (2007), it is possible that participants decide on a likely referent for the pronoun even without disambiguating cues under deep processing. If so, this would have to be re-evaluated at the disambiguation point. If, on the other hand, participants refrain from singling out a possible referent early on, and wait until disambiguating information becomes available (behaviour which may be reinforced throughout the experiment as they discover multiple items in which the pronoun is disambiguated downstream), then Stage two evaluation would take place at the disambiguation point. In either scenario, then, some kind of stage two evaluation takes place at the disambiguation point; this is the effect that we are interested in picking up.

(masculine) so that both matched the pronoun, creating a temporary ambiguity about the referent of the pronoun. The correct referent of the pronoun only became clear at the repeated NP. The position of the correct referent (NP1 or NP2) was manipulated to create the two antecedent conditions. An example of the conditions is given in (4) to (5) below. (Note that the idiomatic translation given at the end of each item set covers all four conditions.)

(4) **Early disambiguation**

4a. *ER-NP1*
Der Trainer hat die Spielerin getroffen, um die Turnschuhe abzugeben. Danach hat er überraschenderweise die Spielerin zur Abschiedsfeier eingeladen.

Der	Trainer	hat	die	Spieler-in	getroffen,
the.M.SG.NOM	*trainer.M*	*have.3SG*	*the.F.SG.ACC*	*player-F*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	er	überraschenderweise	die	Spieler-in	
have.3SG	*he*	*surprisingly*	*the.F.SG.ACC*	*player-F*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

4b. *DER-NP1*
Der Trainer hat die Spielerin getroffen, um die Turnschuhe abzugeben. Danach hat der überraschenderweise die Spielerin zur Abschiedsfeier eingeladen.

Der	Trainer	hat	die	Spieler-in	getroffen,
the.M.SG.NOM	*trainer.M*	*have.3SG*	*the.F.SG.ACC*	*player-F*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	der	überraschenderweise	die	Spieler-in	
have.3SG	*he.DEM*	*surprisingly*	*the.F.SG.ACC*	*player-F*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

4c. *ER-NP2*
Die Spielerin hat den Trainer getroffen, um die Turnschuhe abzugeben. Danach hat er überraschenderweise die Spielerin zur Abschiedsfeier eingeladen.

Die	Spieler-in	hat	den	Trainer	getroffen,
the.F.SG.NOM	*player-F*	*have.3SG*	*the.M.SG.ACC*	*trainer.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	er	überraschenderweise	die	Spieler-in	
have.3SG	*he*	*surprisingly*	*the.F.SG.ACC*	*player-F*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

4d. *DER-NP2*
Die Spielerin hat den Trainer getroffen, um die Turnschuhe abzugeben. Danach hat der überraschenderweise die Spielerin zur Abschiedsfeier eingeladen.

Die	Spieler-in	hat	den	Trainer	getroffen,
the.F.SG.NOM	*player-F*	*have.3SG*	*the.M.SG.ACC*	*trainer.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	der	überraschenderweise	die	Spieler-in	
have.3SG	*he.DEM*	*surprisingly*	*the.F.SG.ACC*	*player-F*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

"The {trainer/player} met the {player/trainer} in order to hand in the plimsolls. Afterwards he surprisingly invited the player to a farewell party."

(5) **Late disambiguation**

5a. *ER-NP1*
Der Trainer hat den Spieler getroffen, um die Turnschuhe abzugeben. Danach hat er überraschenderweise den Spieler zur Abschiedsfeier eingeladen.

Der	Trainer	hat	den	Spieler	getroffen,
the.M.SG.NOM	*trainer.M*	*have.3SG*	*the.M.SG.ACC*	*player.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	er	überraschenderweise	den	Spieler	
have.3SG	*he*	*surprisingly*	*the.M.SG.ACC*	*player.M*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

5b. *DER-NP1*
Der Trainer hat den Spieler getroffen, um die Turnschuhe abzugeben. Danach hat der überraschenderweise den Spieler zur Abschiedsfeier eingeladen.

Der	Trainer	hat	den	Spieler	getroffen,
the.M.SG.NOM	*trainer.M*	*have.3SG*	*the.M.SG.ACC*	*player.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	der	überraschenderweise	den	Spieler	
have.3SG	*he.DEM*	*surprisingly*	*the.M.SG.ACC*	*player.M*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

5c. *ER-NP2*
Der Spieler hat den Trainer getroffen, um die Turnschuhe abzugeben. Danach hat er überraschenderweise den Spieler zur Abschiedsfeier eingeladen.

Der	Spieler	hat	den	Trainer	getroffen,
the.M.SG.NOM	*player.M*	*have.3SG*	*the.M.SG.ACC*	*trainer.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	er	überraschenderweise	den	Spieler	
have.3SG	*he*	*surprisingly*	*the.M.SG.ACC*	*player.M*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

5d. *DER-NP2*
Der Spieler hat den Trainer getroffen, um die Turnschuhe abzugeben. Danach hat der überraschenderweise den Spieler zur Abschiedsfeier eingeladen.

Der	Spieler	hat	den	Trainer	getroffen,
the.M.SG.NOM	*player.M*	*have.3SG*	*the.M.SG.ACC*	*trainer.M*	*meet.PTCP*
um	die	Turnschuh-e	ab<zu>geben.	Danach	
in.order	*the.F.PL.ACC*	*plimsoll.M-PL*	*<to>hand.in.INF*	*afterwards*	
hat	der	überraschenderweise	den	Spieler	
have.3SG	*he.DEM*	*surprisingly*	*the.M.SG.ACC*	*player.M*	
zu-r	Abschiedsfeier	eingeladen.			
to-F.DAT	*farewell.party.F*	*invite.PTCP*			

> "The {trainer/player} met the {player/trainer} in order to hand in the plimsolls. Afterwards he surprisingly invited the player to a farewell party."

The 64 experimental items were interspersed with 96 fillers and divided over 8 lists in a Latin-square design. The number and type of referents in both sentences of the fillers was varied to prevent participants building expectations about the position and type of referents in the experimental items. 72 of the fillers contained feminine pronouns *sie* or *die* ("her").

4.2 Participants

Data was collected from 65 native German speakers (10 male), age range 18-35 years, of whom 54 were included in the analysis based on their accuracy (>75%) in the comprehension questions. No participant reported any language disorders. All participants gave their consent and received a small fee or course credit for participation.

4.3 Procedure

Participants were seated with their eyes 60cm from the computer screen displaying the text, with their head supported by a chin-rest and forehead-rest. Texts were displayed on the screen in a black font on a grey background using the Courier New font. Participants were asked to read sentences silently from computer screen at their normal reading rate and to answer comprehension questions by pressing a button on a keyboard, while their eye movements were recorded using the Eyelink 1000 (SR Research) desktop mount. Comprehension questions (Y/N) followed 42 of the experimental items; 32 questions directly probed the referent of the pronoun. The experiment took around 50 minutes, and participants were paid a small amount or given course credit for their participation.

4.4 Analysis

4.4.1 Regions of Interest

For the analysis, the sentences were divided into regions. The pronoun region always appeared on the second line of text. The analysis was carried out on the pronoun region, which always contained the auxiliary verb and the pronoun, the buffer region which contained the adverbial phrase, and the repeated NP region.

4.4.2 Data cleaning procedure

The data cleaning procedure was the same as in Experiment 1, except that the maximum length for an individual fixation was set to 1000ms. Skipping rates were 12.3% in the pronoun region, 2.9% in the buffer region and 0.6% in the repeated NP region.

4.4.3 Data analysis procedure

The following reading measures were calculated: first-pass times; right-bound reading times; rereading times; and total viewing times (see Experiment 1 for definitions). These reading measures were analysed in linear mixed-effects models per region/measure using the package lmerTest version 2.0.33 (Kuznetsova et al., 2016) in the R statistical program (R Core Team, 2017). Before submitting it to the model the data was transformed. The transformation was determined with the Box-Cox procedure (Box & Cox, 1964) using recommendations from Osborne (2010). The fixed part of the model contained the sum-coded factors pronoun (*der*, *er*); antecedent (*NP1*, *NP2*); ambiguity (*early disambiguation*, *late disambiguation*) and the trial number (centred), as well as the interactions between pronoun, antecedent and ambiguity. The random part of the model contained random intercepts for participant and item. The inclusion of by-item and by-participant random slopes for pronoun and antecedent were determined using the function rePCA from the package RePsychLing (Baayen et al., 2015). Where interactions were followed with pairwise comparisons, *p*-values were Bonferroni-corrected.

4.5 Predictions

Under deep processing conditions (i.e. all conditions in this experiment), it should be easier to process *er* when it refers to the more prominent NP1 (proto-agent), compared to the less prominent NP2 (proto-patient). Conversely, it should be easier to process *der* when it refers to NP2 compared to NP1. This should lead to an interaction of pronoun and antecedent: ER-NP1 will have shorter reading times than ER-NP2 and DER-NP2 will have shorter reading times than DER-NP1. Our interest lies in identifying when such interactions take place.

If prominence information is used evaluatively as soon as a unique antecedent is identified (i.e. at stage two in two-stage processing models), then the timing of the above prominence effects will differ between the early and late disambiguation conditions: In the early disambiguation conditions, where gender cues identify a single antecedent, prominence effects should be visible in earlier measures (first-pass times, right-bound reading times) in the pronoun and buffer regions. In contrast, in the late disambiguation conditions, prominence effects should be delayed until the disambiguating information is read, i.e. in the repeated NP region (early or late measures) and possibly also in later measures in the pronoun and buffer regions.

Additionally, if the demonstrative pronoun *der* exerts more processing demands than the personal pronoun *er*, as seen in Experiment 1, we will see overall higher reading times for *der* compared to *er*: this should be detected in the cumulative measure (total viewing times) and may be visible in either earlier or later measures.

Furthermore, there may be an overall processing cost for late disambiguation conditions compared to early disambiguation conditions, because of the additional strategic engagement needed to deselect one of the potential antecedents (Stewart et al., 2007).

4.6 Results

4.6.1 Comprehension questions

Overall accuracy on the comprehension questions was 89%, SD 6, range 76-100. Accuracy on experimental items was also 89%, SD 6, range 76-100.

4.6.2 Skipping rates

Skipping rates for the critical region (auxiliary + pronoun) were 12.3%; for the buffer region 2.9%; and for the repeated NP region 0.6%.

4.6.3 Eyetracking results

The means for each region and measure are shown in Table 6. Table 7 shows the outcome of the statistical models. Results are described in more detail in the sections below; total viewing times are not discussed, since they simply reflect the effects that are found in the early or the late measures.

		First-pass times	Right-bound reading times	Rereading times	Total viewing times
	PRONOUN REGION				
		Mean	Mean	Mean	Mean
Early disambiguation	*DER-NP1*	262 (164)	334 (189)	426 (296)	488 (305)
	DER-NP2	265 (167)	330 (181)	360 (213)	471 (254)
	ER-NP1	243 (133)	287 (140)	312 (195)	364 (219)
	ER-NP2	244 (128)	299 (150)	353 (263)	394 (247)
Late disambiguation	*DER-NP1*	289 (182)	347 (178)	408 (266)	514 (301)
	DER-NP2	283 (177)	330 (192)	413 (413)	505 (326)
	ER-NP1	243 (121)	288 (146)	315 (203)	380 (218)
	ER-NP2	249 (129)	300 (148)	352 (232)	402 (243)
	BUFFER REGION				
Early disambiguation	*DER-NP1*	342 (225)	413 (259)	468 (389)	562 (393)
	DER-NP2	338 (205)	418 (241)	406 (293)	536 (334)
	ER-NP1	337 (196)	383 (223)	401 (309)	477 (328)
	ER-NP2	323 (190)	369 (225)	482 (447)	509 (391)
Late disambiguation	*DER-NP1*	337 (195)	408 (237)	493 (349)	592 (378)
	DER-NP2	329 (185)	391 (240)	490 (481)	582 (437)
	ER-NP1	328 (210)	358 (219)	420 (291)	511 (335)
	ER-NP2	341 (203)	374 (231)	476 (359)	549 (386)
	REPEATED NP REGION				
Early disambiguation	*DER-NP1*	359 (172)	406 (192)	445 (361)	524 (320)
	DER-NP2	381 (209)	419 (226)	431 (325)	535 (326)
	ER-NP1	360 (171)	381 (191)	378 (341)	479 (303)
	ER-NP2	369 (178)	399 (193)	450 (478)	534 (397)
Late disambiguation	*DER-NP1*	343 (155)	382 (185)	478 (350)	552 (552)
	DER-NP2	351 (185)	399 (212)	509 (446)	599 (430)
	ER-NP1	338 (160)	372 (189)	426 (308)	514 (318)
	ER-NP2	368 (368)	394 (185)	569 (482)	600 (464)

Table 6. Means (in ms) of the first-pass times, right-bound reading times, rereading times and total viewing times for each region (pronoun, buffer, repeated NP) per condition in Experiment 3. Standard deviations shown in parentheses.

	First-pass times		Right-bound reading times		Rereading times		Total viewing times	
Effect	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*	*Estimate (SE)*	*t -value*
PRONOUN REGION								
Pronoun	3.770e-02 (8.826e-03)	4.271***	5.418e-02 (9.430e-03)	5.745***	9.093e-02 (1.621e-02)	5.611***	1.188e-01 (1.079e-02)	11.013***
Antecedent	-1.927e-03 (8.826e-03)	-0.218	-8.169e-04 (7.557e-03)	-0.108	-5.330e-03 (1.504e-02)	-0.354	-9.930e-03 (9.069e-03)	-1.095
Ambiguity	-2.166e-02 (8.830e-03)	-2.453*	-7.241e-03 (7.560e-03)	-0.958	-1.110e-02 (1.501e-02)	-0.739	-2.323e-02 (9.072e-03)	-2.560*
Pronoun x Antecedent	3.586e-03 (8.824e-03)	0.406	1.977e-02 (7.556e-03)	2.617**	3.135e-02 (1.504e-02)	2.084*	1.722e-02 (9.068e-03)	1.899(*)
Pronoun x Ambiguity	-1.365e-02 (8.829e-03)	-1.546	-4.315e-03 (7.560e-03)	-0.571	-4.501e-03 (1.501e-02)	-0.300	-2.851e-03 (9.072e-03)	-0.314
Antecedent x Ambiguity	-1.607e-03 (8.831e-03)	-0.182	-7.636e-03 (7.561e-03)	-1.010	1.437e-02 (1.503e-02)	0.956	-9.435e-03 (9.073e-03)	-1.040
Pronoun x Antecedent x Ambiguity	-2.884e-03 (8.832e-03)	-0.327	-9.151e-03 (7.562e-03)	-1.210	6.598e-03 (1.505e-02)	0.438	3.454e-04 (9.074e-03)	0.038
BUFFER REGION								
Pronoun	1.057e-02 (9.010e-03)	1.173	4.866e-02 (8.140e-03)	5.978***	3.246e-02 (2.348e-02)	1.382	5.866e-02 (1.071e-02)	5.475***
Antecedent	3.037e-03 (7.836e-03)	0.388	3.662e-03 (7.302e-03)	0.501	-2.306e-03 (1.738e-02)	-0.133	-5.138e-03 (8.921e-03)	-0.576
Ambiguity	-3.900e-03 (7.843e-03)	-0.497	1.816e-02 (7.308e-03)	2.485*	-3.487e-02 (1.736e-02)	-2.008*	-3.026e-02 (8.929e-03)	-3.389***
Pronoun x Antecedent	4.203e-03 (7.836e-03)	0.536	-2.314e-04 (7.302e-03)	-0.032	5.360e-02 (1.737e-02)	3.086**	1.871e-02 (8.921e-03)	2.097*
Pronoun x Ambiguity	3.265e-04 (7.852e-03)	0.042	-2.554e-03 (7.316e-03)	-0.349	-1.495e-02 (1.734e-02)	-0.862	-4.873e-04 (8.939e-03)	-0.055
Antecedent x Ambiguity	9.518e-03 (7.845e-03)	1.213	5.272e-03 (7.309e-03)	0.721	-2.966e-03 (1.745e-02)	-0.170	1.818e-03 (8.931e-03)	0.204
Pronoun x Antecedent x Ambiguity	-1.350e-02 (7.845e-03)	-1.721(*)	-1.885e-02 (7.311e-03)	-2.579**	9.377e-03 (1.742e-02)	0.538	-4.709e-03 (8.932e-03)	-0.527
REPEATED NP REGION								
Pronoun	-4.701e-03 (6.821e-03)	-0.689	1.797e-02 (6.304e-03)	2.851**	2.836e-02 (1.788e-02)	1.586	2.314e-02 (9.212e-03)	2.511*

Antecedent	-1.666e-02 (6.817e-03)	-2.444*	-1.904e-02 (6.300e-03)	-3.022**	-3.720e-02 (1.792e-02)	-2.075*	-3.173e-02 (8.127e-03)	-3.904***
Ambiguity	2.318e-02 (6.823e-03)	3.397***	1.862e-02 (6.307e-03)	2.953**	-7.743e-02 (1.793e-02)	-4.318***	-2.759e-02 (8.135e-03)	-3.391***
Pronoun x Antecedent	6.493e-03 (6.817e-03)	0.953	8.691e-03 (6.301e-03)	1.379	4.277e-02 (1.795e-02)	2.383*	1.694e-02 (8.128e-03)	2.084*
Pronoun x Ambiguity	7.937e-03 (6.827e-03)	1.163	9.674e-03 (6.310e-03)	1.533	2.521e-02 (1.788e-02)	1.410	5.829e-03 (8.140e-03)	0.716
Antecedent x Ambiguity	4.094e-04 (6.824e-03)	0.060	4.071e-03 (6.307e-03)	0.645	1.313e-02 (1.790e-02)	0.734	7.581e-03 (8.135e-03)	0.932
Pronoun x Antecedent x Ambiguity	-1.226e-02 (6.822e-03)	-1.797(*)	1.895e-03 (6.305e-03)	0.301	-9.399e-03 (1.794e-02)	-0.524	5.094e-03 (8.133e-03)	0.626

Table 7. Model outputs for all measures (first-pass times, right-bound reading times, rereading times and total viewing times) in the pronoun, buffer and repeated NP regions. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; (*) $p < 0.1$.

4.6.3.1 Early measures, pronoun region

In first-pass times and right-bound reading times in the pronoun region there is a main effect of pronoun, with DER conditions taking longer than ER conditions. In first-pass times there is a main effect of ambiguity, with late disambiguation conditions taking longer than early disambiguation conditions. In the right-bound reading times there is an interaction of pronoun and antecedent: follow-up pairwise comparisons did not reach significance.

4.6.3.2 Early measures, buffer region

In the buffer region the right-bound reading times show a main effect of pronoun, with DER conditions taking longer than ER conditions, and a main effect of ambiguity, with early disambiguation conditions taking longer than late disambiguation conditions. These effects are qualified by a significant interaction of pronoun, antecedent and ambiguity. Follow-up pairwise comparisons did not reach significance.

4.6.3.3 Early measures, repeated NP region

In the repeated NP region the right-bound reading times show a main effect of pronoun, with DER conditions taking longer than ER conditions. There is a main effect of antecedent in both first-pass times and right-bound reading times, with NP2 conditions taking longer than NP1 conditions. There is a main effect of ambiguity in first-pass times and right-bound reading times, with early disambiguation conditions taking longer than late disambiguation conditions. In the first-pass times there is a marginal interaction between pronoun, antecedent and ambiguity. Follow-up pairwise comparisons reveal that in the late disambiguation conditions, first-pass times are significantly shorter when *er* refers to NP1 compared to NP2 (t = -2.912, p = 0.01).

4.6.3.4 Later measures, pronoun region

In the pronoun region there is a main effect of pronoun in rereading times, with *der* taking longer than *er*. In rereading times there is a significant interaction of pronoun and antecedent. Follow-up pairwise comparisons do not reach significance.

4.6.3.5 Later measures, buffer region

There is a main effect of ambiguity in rereading times, with late disambiguation conditions taking longer than early disambiguation conditions. And there is a significant interaction of pronoun and antecedent in rereading times. Follow-up pairwise comparisons reveal that rereading times are significantly longer when *der* refers to NP1 compared to NP2 (t = 2.276, p = 0.05), and that rereading times are marginally shorter when *er* refers to NP1 compared to NP2 (t = -2.123, p = 0.07).

4.6.3.6 Later measures, repeated NP region

There is a main effect of antecedent in the rereading times, with NP2 conditions taking longer than NP1 conditions. There is also a main effect of ambiguity in rereading times, with late disambiguation conditions taking longer than early disambiguation conditions. And finally there is a significant interaction of pronoun and antecedent in the rereading times. Pairwise comparisons reveal that *er* has significantly longer rereading times when referring to NP2 compared to NP1 (rereading t = -3.052, $p < 0.01$).

4.7 Discussion

In this experiment, in which deep processing was encouraged by frequently probing the referent of the pronoun, there are three main findings: increased processing times for the demonstrative

pronoun *der* compared to the personal pronoun *er*; increased processing times when the pronoun was ambiguous; and pronoun by antecedent interactions.

The main effect of pronoun, with increased reading times for the demonstrative, was evident in right-bound reading times in all three regions, and first-pass and rereading times in the pronoun region. The effect here appears not only on the pronoun itself but two regions downstream, in both earlier and later measures. This finding looks very similar to the main effect of pronoun in Experiment 1. There, the effect was attributed to the unexpectedness of a form that signals a referential shift, as opposed to the personal pronoun which signals referential maintenance (regardless of the actual referent). But there are two important differences. Firstly, some of the main effects of pronoun are qualified by an interaction with antecedent. Secondly, the conditions of Experiment 1 gave rise to shallow processing, while in the current experiment, deeper processing was encouraged. So it is possible that the pronoun effect here arises for different reasons than in Experiment 1. Considering again the discourse functions of the demonstrative, previous studies have attributed the increased processing for the demonstrative to the fact that it must retrieve a less prominent, and therefore less accessible, antecedent from the previous discourse, and also the forward function of topic shift for the upcoming discourse (Schumacher et al., 2015, 2017). In the present experiment this would only apply to the *DER-NP2* conditions where the antecedent is less accessible, which could account for cases where there is also an interaction of pronoun and antecedent. This still leaves some main effects of pronoun (without interactions) unexplained. We propose the following: while the *DER-NP1* condition has the advantage that a more prominent antecedent is being retrieved, it has the disadvantage that resolving to NP1 clashes with the requirement that the demonstrative refers to a non-prominent antecedent. In effect, in the *DER-NP1* condition participants are forced to resolve the pronoun to a more prominent, but less felicitous, antecedent. This clash may result induce a processing cost, reflected in longer reading times. This would mean that there is a processing cost associated with both the *DER-NP2* conditions (accessibility) and the *DER-NP1* conditions (infelicity qua the anti-agent preference).

Another finding from Experiment 3 were main effects of ambiguity. It was predicted, based on Stewart et al.'s (2007) study, that ambiguous pronouns (late disambiguation conditions) would take longer than unambiguous pronouns (early disambiguation conditions), because of the longer time required for disengagement during the second processing stage. But the picture here looks to be a little more complex. The predicted effect was found in the first-pass times in the pronoun region, but the pattern is then reversed in early measures at the buffer and repeated NP regions, with early disambiguation conditions taking longer; it reverses back again in the later measure in these two regions, where the late disambiguation conditions take longer. Stewart et al. (2007) attribute the effect to the increased strategic processing required to reduce activation of one of the potential antecedents; in other words, participants have to actively decide between two (or more) potential referents and this requires additional attention. We suggest that this applies when the pronoun is first encountered (reflected in early measures at the pronoun region). But, precisely because a single candidate antecedent has been identified in the early disambiguation conditions, the integration of that antecedent can begin earlier than in late disambiguation conditions, and this integration is reflected in longer processing times in early measures in the buffer region. This may be due to participants reading on more quickly in the late disambiguation conditions as they try to obtain more information, or it may be due to a slowdown in reading the early disambiguation conditions while integration starts. Tentative evidence that integration has already started for early disambiguation conditions at that point is that the pronoun by antecedent interaction differs over ambiguity (resulting in a pronoun by antecedent by ambiguity interaction at that point). However, it is difficult to be sure about what is happening here because pairwise differences are not significant. At this stage, there would be no need to do additional processing for the ambiguous pronoun because the relevant disambiguating information is not yet available. When it does become available at the repeated NP region, there is a (later) slowdown for the late

disambiguation conditions while the antecedent is integrated, resulting in the main effect of ambiguity in that region.

There is strong evidence that prominence information affects pronoun resolution during deep processing. There are a number of pronoun by antecedent interactions, and in particular the reading times for the personal pronoun are slower when it refers to a less prominent antecedent. The trend is for reading times of the demonstrative to be slower when it refers to a more prominent antecedent, although the pairwise differences are significant only in one measure. These effects confirm findings from earlier studies, that the prominence information affects the personal and demonstrative pronouns differently (Bosch et al., 2007; Schumacher et al., 2015, 2016, 2017).

With regard to the timing of the prominence information, the picture is more mixed. We expected that, under deep processing, the use of the prominence information in evaluating pronoun-antecedent links (i.e. at Stage 2 of the two-stage model) would lead to antecedent by pronoun interactions. Furthermore, we expected that the interactions would occur earlier in the early disambiguation conditions than in the late disambiguation conditions, because we assumed that Stage 2 would begin when disambiguating information became available (for the early disambiguation conditions, this is at the pronoun region and for the late disambiguation conditions this is at the repeated NP region). There is limited evidence for this timing distinction in right-bound reading times in buffer region, where there is a three-way interaction of pronoun, antecedent and ambiguity, indicating that the pronoun by antecedent interaction differs between early and late disambiguation conditions; however, the pairwise comparisons were not significant. At the repeated NP region, first-pass times reveal a prominence effect in the expected direction for the personal pronoun in the late disambiguation conditions. This could indicate that the evaluative use of prominence information may be delayed until disambiguating information becomes available, but once it does become available, prominence information is deployed rapidly. On the other hand, there is also evidence for prominence information affecting resolution at some distance from the point at which a single candidate antecedent could be identified: later measures in the buffer and repeated NP region show pronoun by antecedent interactions that do not differ between ambiguous and unambiguous conditions, indicating that prominence information is affecting both early and late disambiguating conditions similarly. For the early disambiguation conditions, this is at some distance from the point at which a single candidate could be identified. It is possible that the availability of the repeated NP in both early and late disambiguation conditions (effectively creating a second unambiguous cue to pronoun resolution in the early disambiguation conditions) encouraged participants to delay resolving the pronoun until the repeated NP had been encountered in some cases. This would add further weight to the argument that prominence information is deployed strategically during pronoun resolution.

A final observation from this experiment is that when prominence information was used, more differences were found for the personal than the demonstrative pronoun. Differences for the demonstrative pronoun may have been masked by overall increased reading times in those conditions. It was suggested above that, for the processing of the demonstrative pronoun, both the position on the prominence hierarchy (i.e. accessibility) and the felicity of referring to a particular position on the hierarchy may have had an effect on processing times. This fits with the finding of more differences for the personal pronoun compared to the demonstrative. For the personal pronoun, the ER-NP1 condition is advantageous for both felicity and accessibility, and the ER-NP2 condition is disadvantageous for both. If felicity and accessibility are combined, the difference between ER-NP1 and ER-NP2 is larger than between DER-NP1 and DER-NP2.

5 General discussion

We conducted three experiments investigating how prominence information is used during the resolution of gender-disambiguating German personal and demonstrative pronouns.

Demonstrative pronouns are particularly interesting because, unlike personal pronouns, they tend to be resolved to antecedents that are not the most prominent entities in the prior discourse. This is a challenge for processing models which are built on the assumption that pronouns always seek prominent antecedents. In comparison to many previous studies, we presented unambiguous pronouns while still manipulating prominence.

Experiment 1 was an eye-tracking during reading experiment in which gender features of the pronoun and antecedent made the pronouns unambiguous. The original assumption was that the pronouns would be fully resolved, enabling us to explore the prominence hierarchies of thematic and grammatical roles during the resolution of personal and demonstrative pronouns. But, while demonstrative pronouns elicited higher reading times than personal pronouns throughout the experiment, we found that the prominence of the antecedents did not make any difference to participants' reading behaviour. In Experiment 1 the comprehension questions rarely probe the referent of the pronoun, creating conditions under which participants may have engaged in shallow processing. During shallow processing, pronouns may not be fully resolved. If this was indeed the case in Experiment 1, as we argue below, it would suggest that prominence information is not part of the automatic co-indexation process that has been claimed to form the first part of pronoun resolution (Rigalleau et al., 2004; Stewart et al., 2007). However, this seeming discrepancy may arise from differing concepts of prominence and how prominence information is used. We address this point below.

Given that gender cues in Experiment 1 made available only one potential referent from the context sentence, the increased reading times for the demonstrative pronoun under shallow conditions remain surprising. We find an explanation of low-level factors such as frequency and ambiguity unconvincing given the stability of the effect over the course of the experiment, and we suggest instead that there is a cost in encountering a form which signals referential shift rather than topic maintenance, regardless of whether a single referent is eventually identified.

Experiment 2 was a rating task using a subset of materials from Experiment 1. The task involved participants having to reflect on each sentence in turn and make a judgment about it, and as such encouraged a deeper level of processing. Even though participants were not explicitly asked to make judgments about the referents of the pronouns, the results showed that the prominence information did affect participants' rating of the items; specifically, the items containing demonstrative pronouns received lower ratings when the pronoun referred to a prominent antecedent. This is in line with previous research showing that demonstrative pronouns are preferentially resolved to non-prominent antecedents, and that the personal pronoun is somewhat more flexible in its interpretation preferences (Bosch & Umbach, 2007; Schumacher et al., 2016; for similar findings with Dutch demonstratives, see Kaiser, 2011b). The finding that prominence information affected ratings despite the presence of gender cues rules out the possibility that gender cues block the use of weaker resolution cues such as prominence, as long as the task encourages deep processing. Furthermore, the differing role of prominence information in Experiments 1 and 2, which differ in task and depth of processing, suggests that the evaluative use of prominence information is under strategic control during pronoun resolution.

In Experiment 3, an eye-tracking during reading experiment, we created conditions for deep processing by increasing the number of comprehension questions that directly probed the referent of the pronoun. Under these conditions we manipulated pronoun type, reference to a more or less prominent antecedent, and the ambiguity of the pronoun in order to assess the timecourse of processing associated with prominence information. As in Experiment 1, demonstrative pronouns elicited longer reading times than personal pronouns. In addition, there were interactions of pronoun and antecedent which suggested that prominence information was being used during resolution. This demonstrates that under deeper processing conditions prominence information has an effect that is not captured under shallow processing, and is therefore likely to be under strategic control. The precise timing of the prominence information

was less clear, however. Interactions with ambiguity suggested that prominence information was (sometimes) deployed rapidly as soon as a single candidate antecedent could be identified. But there was also evidence that the use of the prominence information was somewhat delayed. It could be the case that participants differed in how rapidly they deployed the prominence information, or that individual participants changed their strategy over the course of the experiment.

In sum, we would like to argue that shallow processing was involved in Experiment 1, but not Experiments 2 and 3, and on this premise we claim that the combined results of the three experiments together demonstrate that the evaluative use of prominence information during pronoun resolution is under strategic control. In order to substantiate this claim, it is therefore important to first establish that in Experiment 1, unlike Experiments 2 and 3, shallow processing was involved. After conducting Experiment 1 we considered two alternative explanations for not finding the expected interaction of pronoun and antecedent. Firstly, it could be the case that the particular materials we used did not give rise to the expected prominence effect. However, given that Experiment 2 used a subset of the same materials and did show the effect, this possibility can be dismissed. A second possibility we considered was that only ambiguous pronouns show prominence effects, and that the presence of disambiguating gender features overrules the need for considering prominence. This possibility was again ruled out by the results of Experiment 2, which showed prominence effects despite presenting gender disambiguated pronouns. Finally, we consider task effects. Experiment 2, unlike Experiment 1, required participants to make a meta-linguistic judgment about each item. Such an evaluative task is very likely to enhance participant engagement with the presented material. We would further argue that our experimental set up for Experiment 1, where comprehension questions rarely targeted the reference of the pronoun, were precisely the conditions under which shallow processing was shown to take place in Stewart et al. (2007), and we followed their set-up using more frequent, targeted comprehension questions to enhance the depth of processing for Experiment 3.

Within the framework of a two-stage model of pronoun resolution such as put forward by Rigalleau et al. (2004) and elaborated by Stewart et al. (2007), the first, automatic stage of pronoun resolution proceeds whether or not processing is shallow or deep. This stage involves indexing potential antecedents using gender features. Under shallow processing conditions, the process stops at this point. Under deeper processing conditions, resolution proceeds to Stage 2, which involves strategic processing. Here, links between the pronoun and the candidate antecedents are evaluated and the activation of the non-antecedent is reduced. We propose that prominence information is used at Stage 2 to evaluate candidate antecedents. As such, prominence information may not be deployed in shallow processing, where pronoun resolution is stopped after the first, automatic stage.

The question of whether prominence information is used at Stage 1, Stage 2 or both may well depend on the precise notion of prominence information and how that information is used. Rigalleau et al. (2004) invoke the notion of accessibility in order to determine which antecedents get checked (for gender cues) at Stage 1 and which do not. They follow Greene et al. (1992) in assuming that the available antecedents are in the "focus of attention". In their own experiments, Rigalleau and colleagues create scenarios where two potential referents are available in the preceding text, with one at quite some distance from the pronoun (approx. 35 words) and one much closer (6 words); they assume that the close but not the distant referent is accessible. They also imply (p. 919) that appearing in the previous sentence results in enough activation to warrant being the focus of attention. As such, some kind of prominence information must feed into Stage 1, such that certain antecedents are checked and others are not, but this could equate to a simple distance metric or labelling all antecedents accessible if they appear in the preceding sentence. Stewart et al. (2007) similarly assume that both referents in a typical agent–patient or subject–object configuration are potentially indexable (both having appeared in the previous sentence). In a scenario such as ours in Experiment 1, then, where both potential antecedents appear in the

preceding sentence and where the less prominent antecedent is linearly closer to the pronoun, it is hard to argue that only the most prominent/accessible referent (NP1) is indexed; it seems more likely that both are indexed. If both are indeed indexed at Stage 1, this means that the relative prominence between an agent and a patient in the preceding sentence, which has been shown to have robust influence on the eventual pronoun resolution preferences of personal and demonstrative pronouns, must be irrelevant at Stage 1, under the model that Rigalleau, Stewart and colleagues propose. We therefore suggest that the type of prominence information that we are interested in, i.e. the relative prominence of an agent compared to a patient, is deployed at Stage 2, under deep processing conditions. We should also emphasise that we are talking about the evaluative use of this prominence information in unambiguous resolution, i.e. when gender features only match one antecedent. For example, a demonstrative pronoun refers to the more prominent antecedent; while the reference is clear, referring to a less prominent antecedent may not be considered felicitous and it is this evaluative use of prominence information we are interested in. As pointed out by a reviewer, placing prominence at Stage 2 would appear to contradict previous literature which claims that prominence or accessibility information is considered early, notably Arnold et al. (2000). (Earlier studies claiming an important role for prominence (e.g. Hudson d'Zmura & Tanenhaus, 1998; Gordon et al., 1993) do not deploy methodologies with a timecourse fine-grained enough to detect the differences that we are interested in). Arnold and colleagues' results, however, do show a difference between ambiguous and unambiguous cases. In the unambiguous conditions, there are no differences between referring to a more or less prominent antecedent in the time-window they examined, which would be expected if prominence were being used to evaluate the candidate antecedent. So while their results certainly reflect the authors' main claim that gender information is used early, it is not clear that their results wholly contradict our claim that evaluative use of prominence takes place at Stage 2. Furthermore, most of the studies showing effects of prominence/accessibility on pronoun resolution have also involved some kind of judgment or evaluation task, which can trigger deep processing and therefore involve Stage 2. Many studies contained offline experiments involving referent selection or sentence continuation tasks, or a judgment task (Bosch et al., 2007; Bouma & Hopp, 2006; 2007; Colonna et al., 2012; Crawley & Stevenson, 1990; Kaiser, 2011a; Kaiser & Trueswell, 2008; Schumacher et al., 2015; 2016; Stevenson et al., 1994). Self-paced reading experiments were combined with referent selection tasks (Bosch et al., 2007; Crawley et al., 1990; Hudson d'Zmura & Tanenhaus, 1998; in Gordon et al., 1993 SPR is combined with a true/false task that is not described in detail). Visual world experiments normally involved a picture-matching or picture judgment task (Arnold et al., 2000; Kaiser & Trueswell, 2008; Kaiser, 2011b; exceptions are Järvikivi et al., 2005 and Schumacher et al., 2017 in which the accompanying continuation task appeared infrequently, rather than after every trial). In the EEG experiment reported in Schumacher et al., 2015, a comprehension question was presented after every trial, and some questions required pronoun resolution.

One further, persistent finding was that the demonstrative pronoun was more costly to process than the personal pronoun. As discussed in Experiment 1, we originally expected to find that the demonstrative pronoun would have a processing cost relative to the personal pronoun because of its distinct discourse function, following Schumacher et al. (2015). The requirement for the demonstrative pronoun to pick out a non-prominent referent as its antecedent is associated with higher processing demands; the demonstrative also has a forward-looking topic shift function which is also associated with higher costs. But both of these discourse functions pertain to the demonstrative referring to the non-prominent antecedent. In Experiment 1, reference to NP1 and NP2 was systematically manipulated. The discourse functions would only explain the processing cost for the demonstrative if there was also an interaction between pronoun and antecedent, which we did not find. Our alternative explanation, then, is that the form itself, which signals a referential shift instead of thematic maintenance, was unexpected and therefore increased processing times, regardless of the eventual referent.

As also discussed in Experiment 1, while there are several word-level factors such as frequency and ambiguity that make the demonstrative pronoun more effortful to process than the personal pronoun, these factors alone seem insufficient to explain why the effect for the demonstrative pronoun does not diminish during the experiments as participants become familiar with the local experimental context. We suggested in Experiment 3, when there was deeper processing, that the cost for the demonstrative may have had a different source, arising from the combination of two discourse factors, accessibility and felicity. When the demonstrative refers to a less prominent antecedent, there is a cost of retrieving a less accessible entity. When it refers to the more accessible antecedent, there is a cost associated with retrieving a less preferred antecedent for the demonstrative. Thus both conditions for the demonstrative pronoun are associated with a processing penalty. The combination of accessibility and felicity can explain why prominence effects for the personal pronoun were easier to detect in Experiment 3 than prominence effects for the demonstrative. Note that this fits with the finding from Experiment 2 that ratings for the demonstrative were affected by prominence information. In the rating task, the cost of retrieving a more or less accessible antecedent would not be visible, since it is not a task that can measure processing. But it does measure felicity: there is a cost when the demonstrative refers to an infelicitous antecedent. The personal pronoun, being more flexible in its referent, was not penalised to the same extent for referring to a dispreferred antecedent.

The relevance of prominence information to pronoun resolution processes is particularly important for languages such as German where the distinction between different pronoun types may be largely dependent on their resolution preferences. Our results align with previous findings that demonstratives require additional processing resources compared to personal pronouns, and that personal pronouns preferentially refer to more prominent antecedents while demonstratives avoid prominent antecedents. Further, we have shown that prominence information is not ruled out by the presence of stronger resolution cues such as gender. However, the deployment of prominence information in the evaluation of candidate antecedents is under strategic control and therefore does not take place under shallow processing conditions.

Acknowledgements

We are grateful to Simon Napierala, Daniela Mertzen, Hilde Penner and Julia Plechatsch for assistance with materials and experimental set-up. To Janna Drummer, Claudia Kilter, Daniela Mertzen, Jana Mewe and Brita Rietdorf for data collection. Thanks to Franziska Kretzschmar for analysis advice. Claudia Felser not only generously provided laboratory space but also offered advice and assistance in the planning stages in adapting the experimental design to an eye-tracking during reading paradigm, and advice on conceptual issues. Finally, we gratefully acknowledge that this research was funded by the German Research Foundation (DFG) as part of the Collaborative Research Center 1252 “Prominence in Language” – Project-ID 281511265 – in the project C07 “Forward and backward functions of discourse anaphora” at the University of Cologne, Department of German Language and Literature I, Linguistics.

References

Ahrenholz, B. (2007). Verweise mit Demonstrativa im gesprochenen Deutsch: Grammatik, Zweitspracherwerb und Deutsch als Fremdsprache. Berlin: De Gruyter.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, *106*(4), 748–765.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge.

Arnold, J. (1998). *Reference form and discourse patterns* (PhD Thesis). Stanford University.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13–B26. http://dx.doi.org/10.1016/S0010-0277(00)00073-1

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 748–769.

Bosch, P., & Hinterwimmer, S. (2016). Anaphoric reference by demonstrative pronouns in German. In *Empirical Perspectives on Anaphora Resolution* (pp. 193–212). De Gruyter.

Bosch, P., Katz, G., & Umbach, C. (2007). The non-subject bias of German demonstrative pronouns. In *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference* (pp. 145–164).

Bosch, P., Rozario, T., & Zhao, Y. (2003). Demonstrative pronouns and personal pronouns: German der versus er. In *Proceedings of the EACL 2003*. Budapest.

Bosch, P., & Umbach, C. (2007). Reference determination for demonstrative pronouns. In D. Bittner (Ed.), *Proceedings of Conference on Intersentential Pronominal Reference in Child and Adult Language*. Berlin.

Bouma, G., & Hopp, H. (2006). Effects of word order and grammatical function on pronoun resolution in German. In *Ambiguity in Anaphora* (pp. 5–13).

Bouma, G., & Hopp, H. (2007). Coreference preferences for personal pronouns in German. In D. Bittner & N. Gagarina (Eds.), *Intersentential pronominal reference in child and adult language* (pp. 53–74). Berlin.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: interpretation of personal and demonstrative pronouns. *Journal of Memory and Language, 53*, 292–313. https://doi.org/10.1016/j.jml.2005.03.003

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159–168. https://doi.org/10.1016/j.bandl.2006.04.005

Clark, H., & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, *7*(1), 35–41. https://doi.org/10.3758/BF03196932

Çokal, D., Sturt, P., & Ferreira, F. (2018). Processing of *it* and *this* in written narrative discourse. *Discourse Processes*, *55*(3), 272–289. https://doi.org/10.1080/0163853X.2016.1236231

Colonna, S., Schimke, S., & Hemforth, B. (2012). Information structure effects on anaphora resolution in German and French: A crosslinguistic study of pronoun resolution. *Linguistics*, *50*(5), 991–1013.

Crawley, R. A., & Stevenson, R. J. (1990). Reference in single sentences and in texts. *Journal of Psycholinguistic Research*, *19*(3), 191–210. https://doi.org/10.1007/BF01077416

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*(4), 245–264. https://doi.org/10.1007/BF01077259

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, *67*(3), 547–619.

Ehrlich, K. (1980). Comprehension of Pronouns. *Quarterly Journal of Experimental Psychology*, *32*(2), 247–255. https://doi.org/10.1080/14640748008401161

Felser, C., & Cunnings, I. (2012). Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics*, *33*(3), 571–603. https://doi.org/10.1017/S0142716411000488

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Patson, N. D. (n.d.). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Filiaci, F., Sorace, A., & Carreiras, M. (2014). Anaphoric biases of null and overt subjects in Italian and Spanish: a cross-linguistic comparison. *Language, Cognition and Neuroscience*, *29*(7), 825–843. https://doi.org/10.1080/01690965.2013.801502

Garnham, A., & Oakhill, J. (1985). On-line resolution of anaphoric pronouns: Effects of inference making and verb semantics. *British Journal of Psychology*, *76*(3), 385–393. https://doi.org/10.1111/j.2044-8295.1985.tb01961.x

Garnham, A., Oakhill, J., & Cruttenden, H. (1992). The role of implicit causality and gender cue in the interpretation of pronouns. *Language and Cognitive Processes*, *7*(3–4), 231–255. https://doi.org/10.1080/01690969208409386

Garrod, S., & Sanford, A. J. (1990). Referential processes in reading: Focusing on roles and individuals. In *Comprehension processes in reading* (pp. 465–485). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Garrod, Simon, & Terras, M. (2000). The Contribution of Lexical and Situational Knowledge to Resolving Discourse Roles: Bonding and Resolution. *Journal of Memory and Language*, *42*(4), 526–544. https://doi.org/10.1006/jmla.1999.2694

Garvey, C., Caramazza, A., & Yates, J. (1974). Factors influencing assignment of pronoun antecedents. *Cognition, 3(3)*, 227–243. https://doi.org/10.1016/0010-0277(74)90010-9

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: the advantage of first mention. *Journal of Memory and Language*, *27*(6), 699–717.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, Names, and the Centering of Attention in Discourse. *Cognitive Science*, *17*(3), 311–347. https://doi.org/10.1207/s15516709cog1703_1

Greene, S. B., McKoon, G., & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 266–283.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307.

Baayen, H., Bates, D., Kliegl, R., & Vasishth, S. (2015). *RePsychLing: Data sets from Psychology and Linguistics experiments*.

Himmelmann, N. P. (1997). *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Hinterwimmer, S. (2015). A unified account of the properties of German demonstrative pronouns. In P. Grosz, P. Patel-Grosz, & I. Yanovich (Eds.), *The Proceedings of the Workshop on Pronominal Semantics at NELS 40* (pp. 61–107). GLSA Publications.

Hirotani, M., & Schumacher, P. B. (2011). Context and topic marking affect distinct processes during discourse comprehension in Japanese. *Journal of Neurolinguistics*, *24*(3), 276–292. http://dx.doi.org/10.1016/j.jneuroling.2010.09.007

Hudson D'Zmura, S., & Tanenhaus, M. K. (1998). Assigning antecedents to ambiguous pronouns: the role of the Center of Attention as the default assignment. In M. A. Walker, A. K. Joshi, & E. F. Prince (eds). *Centering Theory in Discourse*, 199–226. Oxford: Oxford University Press.

Hung, Y.-C., & Schumacher, P. B. (2014). Animacy matters: ERP evidence for the multi-dimensionality of topic-worthiness in Chinese. *Brain Research*, *1555*, 36–47. https://doi.org/10.1016/j.brainres.2014.01.046

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous Pronoun Resolution: Contrasting the First-Mention and Subject-Preference Accounts. *Psychological Science*, *16*(4), 260–264. https://doi.org/10.1111/j.0956-7976.2005.01525.x

Kaiser, E. (2011a). On the relation between coherence relations and anaphoric demonstratives in German. In I. Reich, E. Horch & D. Pauly (eds), *Proceedings of Sinn & Bedeutung 15*, 337–351. Saarbrücken, Germany: Saarland University Press.

Kaiser, E. (2011b). Salience and contrast effects in reference resolution: the interpretation of Dutch pronouns and demonstratives. *Language and Cognitive Processes 26*(10), 1587–1624. http://dx.doi.org/10.1080/01690965.2010.522915

Kaiser, E., & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709–748. https://dx.doi.org/10.1080/01690960701771220

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models*. Retrieved from https://CRAN.R-project.org/package=lmerTest

Love, J., & McKoon, G. (2011). Rules of engagement: incomplete and complete pronoun resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 874–887.

Mcdonald, J. L., & Macwhinney, B. (1995). The Time Course of Anaphor Resolution: Effects of Implicit Verb Causality and Gender. *Journal of Memory and Language*, *34*(4), 543–566. https://doi.org/10.1006/jmla.1995.1025

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098

Osborne, J. W. (2010). Improving your data transformations: applying the Box-Cox transformation. *Practical Assessment, Research and Evaluation*, *15*(12).

Portele, Y., & Bader, M. (2016). Accessibility and Referential Choice: Personal Pronouns and D-pronouns in Written German. *Discours. Revue de Linguistique, Psycholinguistique et Informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics*, *18*, 1–39.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New Arguments in Favour of an Automatic Gender Pronominal Process. *The Quarterly Journal of Experimental Psychology Section A*, *57*(5), 893–933. https://doi.org/10.1080/02724980343000549

Sanford, A.J., Garrod, S., Lucas, A., & Henderson, R. (1983). Pronouns Without Explicit Antecedents? *Journal of Semantics*, *2*(3–4), 303–318. https://doi.org/10.1093/semant/2.3-4.303

Sanford, Anthony J., & Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386. https://doi.org/10.1016/s1364-6613(02)01958-7

Sauermann, A., & Gagarina, N. (2017). Grammatical role parallelism influences ambiguous pronoun resolution in German. *Frontiers in Psychology, 8*, 1205. https://doi.org/10.3389/fpsyg.2017.01205

Schumacher, P. B., Backhaus, J., & Dangl, M. (2015). Backward- and Forward-Looking Potential of Anaphors. *Frontiers in Psychology*, *6*, 1746. https://doi.org/10.3389/fpsyg.2015.01746

Schumacher, P. B., Dangl, M., & Uzun, E. (2016). Thematic role as prominence cue during pronoun resolution in German. In *Empirical Perspectives on Anaphora Resolution* (pp. 213–240). Berlin, Boston: De Gruyter.

Schumacher, P. B., Roberts, L., & Järvikivi, J. (2017). Agentivity drives real-time pronoun resolution: Evidence from German er and der. *Lingua*, *185*, 25–41. http://dx.doi.org/10.1016/j.lingua.2016.07.004

Schütze, C. T., & Sprouse, J. (2014). Judgment data. *Research Methods in Linguistics*, 27–50.
Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*(4), 519–548. https://doi.org/10.1080/01690969408402130
Stewart, A. J., Holler, J., & Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *The Quarterly Journal of Experimental Psychology*, *60*(12), 1680–1696. https://doi.org/10.1080/17470210601160807
Stirling, L., & Huddleston, R. (2002). Deixis and anaphora. In R. Huddleston & G. Pullum (Authors), *The Cambridge Grammar of the English Language* (pp. 1449-1564). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316423530.018
Streb, J., Hennighausen, E., & Rösler, F. (2004). Different anaphoric expressions are investigated by event-related potentials. *Journal of Psycholinguistic Research*, *33*, 175–201.
Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562.
Wang, L., & Schumacher, P. (2013). New is not always costly: evidence from online processing of topic and contrast in Japanese. *Frontiers in Psychology*, *4*, 363. https://doi.org/10.3389/fpsyg.2013.00363
Wilson, F. (2009). *Processing at the syntax–discourse interface in second language acquisition*. Doctoral Thesis, University of Edinburgh.

Appendix A

Table A1. Model outputs for trial by pronoun interaction in all measures of the pronoun and spillover region for active accusative and dative experiencer verbs, Experiment 1.

	First-pass times		Right-bound reading times		Rereading times		Total viewing times	
Effect	*Estimate (SE)*	*t-value*	*Estimate (SE)*	*t-value*	*Estimate (SE)*	*t-value*	*Estimate (SE)*	*t-value*
Pronoun region, active accusatives	-1.468e-06 (5.414e-06)	-0.271	-2.548e-06 (5.499e-06)	-0.463	4.808e-06 (1.493e-05)	0.322	-2.577e-06 (6.002e-06)	-0.429
Spillover region, active accusatives	8.257e-05 (2.354e-04)	0.351	-1.186e-05 (2.250e-04)	-0.053	-1.333e-03 (5.082e-04)	-2.622**	-3.502e-04 (2.544e-04)	-1.376
Pronoun region, dative experiencers	-9.095e-07 (5.009e-06)	-0.182	-2.295e-06 (4.913e-06)	-0.467	1.965e-05 (1.252e-05)	1.569	6.140e-06 (5.541e-06)	1.108
Spillover region, dative experiencers	7.414e-05 (2.046e-04)	0.362	-2.228e-04 (1.947e-04)	-1.144	-8.129e-05 (4.431e-04)	-0.183	-1.174e-04 (2.131e-04)	-0.551