Dialogue & Discourse 17(1) (2026) 89–126 doi: 10.5210/dad.2026.103

It matters how you combine your clauses: Effects of syntactic subordination, connectives, and typographic and prosodic boundaries on the prominence of referents

Timo Buchholz* TIMO.BUCHHOLZ@UNI-TUEBINGEN.DE
University of Tübingen, Germany

Jet Hoek* JET.HOEK@RU.NL
Radboud University Nijmegen, Centre for Language Studies, The Netherlands

Klaus von Heusinger KLAUS.VONHEUSINGER@UNI-KOELN.DE
University of Cologne, Germany

*** Joint first authors**

Editor: Amir Zeldes

Submitted 08/2024; Accepted 02/2026; Published online 02/2026

Abstract

Recency affects how accessible referents are, but the effect of recency is mediated by the structure of the discourse. In a series of four pronoun resolution experiments, we examine how the accessibility of referents is impacted by the form of subsequent discourse segments, investigating effects of syntactic subordination, the presence of explicit coherence markers, and typographic and prosodic boundaries. Our findings indicate that syntactic subordination, connectives, and typographic boundaries all additively contribute to whether an intervening clause is perceived as less or more conceptually integrated, and that this affects how strongly that clause blocks access to a preceding referent. However, the type of prosodic boundary was found to interact with syntax in an unforeseen way: only with syntactic subordination did a high boundary seem to increase the perception of the intervening clause as integrated, but not with coordination. Our results speak to the question of how the mental representation of a discourse is affected by the specific form of the discourse, and call for a reconsideration of intonational boundaries as integratedness cues.

Keywords: coreference, recency, syntactic subordination, connectives, prosody, Right Frontier Constraint

1. Introduction

In a discourse, not all information is equally important, and elements that are important at one point, may not remain that way as the discourse continues. A story can, for example, focus on a single referent for a while, but then introduce a new referent and zoom in on them – in the meantime, the first referent can remain in the picture or can be dropped altogether. The dynamic nature of discourse requires that language users keep track of referents, which includes keeping track of referents' accessibility and relative prominence. This information is for instance vital for the production and interpretation of referring expressions: While highly prominent discourse referents are available to

be picked up by reduced referring expressions, less prominent referents can only be referred to using full descriptive terms (see von Heusinger and Schumacher 2019 for an elaborate discussion).

The prominence of referring expressions seems to be affected by many different factors, including thematic role (e.g., Arnold, 2001; Stevenson et al., 1994), grammatical role (e.g., Arnold, 2001; Fukumura and van Gompel, 2010; Stevenson et al., 1994), topicality (e.g., Cowles and Ferreira, 2012), animacy (e.g., Fukumura and van Gompel, 2011; Vogels et al., 2014), and definiteness (e.g., Brocher and von Heusinger, 2018) . Another factor that appears to affect the prominence of referents is recency: It has long been accepted that a referent has to be at least somewhat recently mentioned in the discourse to be able to be picked up by a reduced referring expression. Indeed, there is evidence from interpretation studies that resolving a pronoun to a recently mentioned referent is easier than resolving a pronoun to a referent whose last mention was further back in the discourse (Ariel, 1990; Clark and Sengul, 1979; Cunnings et al., 2014; Streb et al., 2004). The effect of recency on the prominence of referents, however, is not completely straightforward. Recency can for instance interact with other factors affecting a referent's accessibility mentioned above: Clifton & Ferreira's (1987) findings for instance suggest that recent non-topics are less accessible than non-recent topics. In addition, when talking about recency, it is not obvious in which units recency should be measured: words, clauses, sentences, time (e.g., seconds), the number of other, later-mentioned referents, etc.? Considering working memory constraints on human language processing, linear units of measurements such as the number of intervening words or time are likely to impact the prominence of referents, with referents' prominence level decreasing as more words are uttered and more time elapses after their last mention. However, there are also indications that such a 'simple' measure of recency is insufficient, since it seems that not all linguistic content intervening between a referring expression and its antecedent is created equal. For instance, there is evidence from language production that the number of referents present in the discourse affects choice of referring expression, with fewer reduced forms being produced to refer to referents when other referents have been introduced between the referring expression and its antecedent (Arnold and Griffin, 2007). In addition, a major factor that appears to influence the extent to which intervening content impacts the accessibility of referents is the discourse-structural configuration.

Discourse can be depicted as a hierarchical structure (e.g., Asher, 1993; Asher and Lascarides, 2003; Grosz et al., 1995; Mann and Thompson, 1988; Polanyi, 1988), in which discourse segments (often minimally a grammatical clause, but see Hoek et al., 2018, for a discussion) are in either coordinating or subordinating relations with each other. Asher and Lascarides (2003) illustrate this using the discourse in (1), depicted schematically in Figure 1.

(1) a. John had a great evening last night.
    b. He had a great meal.
    c. He ate salmon.
    d. He devoured lots of cheese.
    e. He then won a dancing competition.

The Elaboration relations in Figure 1 are (discourse-structurally) subordinating; the Narration relations are coordinating. Polanyi (1988) poses that discourse segments are accessible when they are on the 'right frontier' (RF) of a discourse: when there are no subsequent coordinating discourse segments at the same level or a higher level in the discourse structure (see also Asher, 1993; Asher and Lascarides, 2003; Webber, 1988). To illustrate: before the final discourse segment *e* in (1), segments *a*, *b*, and *d* are at the RF, but after *e*, only *a* and *e* are. The Right Frontier Constraint (RFC) predicts

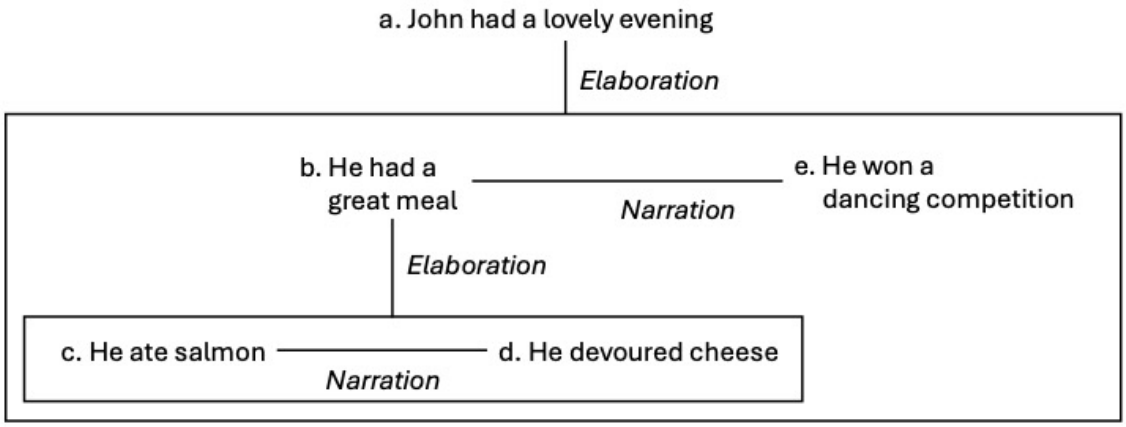

Figure 1: Schematic depiction of the discourse in (1) (Asher and Lascarides, 2003)

that referents introduced in a non-recent discourse segment at the RF are more accessible than referents introduced in a recent discourse segment that is not at the RF. This prediction is borne out in the pronoun interpretation experiment conducted by Holler and Irmen (2007). A discourse-structurally coordinated discourse segment thus impacts the prominence of a preceding referent more than a discourse-structurally subordinated discourse segment: While discourse-structurally subordinated discourse segments might make a preceding referent less prominent by making it less recent, a coordinated discourse segment appears to make a referent unavailable to be picked up by a reduced referring expression.

However, even when an intervening discourse segment (or multiple ones) leaves an earlier discourse referent available to be picked up by a pronoun (because it is discourse-structurally subordinated), an earlier referent's prominence level can vary. One of the additional factors for which there is tentative evidence that it impacts the prominence of referents is the syntactic configuration of the intervening discourse. In a sentence-completion experiment, Cooreman and Sanford (1996), for instance, investigate coreference rates of ambiguous pronouns in short discourses like (2). They find that the pronouns are overwhelmingly (79.8%-92.9%) resolved to the subject of the preceding main clause (the conductor), even if that clause is not the most recent one, like in (2). Both the main and the subordinate clause are at the RF here, so taking only discourse structure into account, we should expect about equal resolution to the two subjects, or even more resolution to the more recent referent. However, Cooreman and Sanford's (1996) results suggest that intervening subordinate clauses may affect the prominence of referents less than intervening main clauses.

(2) The conductor sneezed three times after the tenor opened his music score. He...

Cooreman & Sanford's (1996) findings are in line with the tentative conclusion Clifton and Ferreira (1987) draw on the basis of a post-hoc analysis of their experimental data, where they suggest that resolving a pronoun to a distant topic is easier when intervening referents have been mentioned in subordinate clauses than when they have been mentioned in intervening main clauses.

Why would this be the case? Matthiessen and Thompson (1988) argue that hypotaxis is a grammaticalization of discourse prominence. A key feature of Rhetorical structure theory (RST: Mann

and Thompson 1988) is the Nuclearity principle, which depicts the relation between two discourse segments as either multinuclear, in which both segments are equally important, or mononuclear, in which one segment, the nucleus, is more important, or more central, to the discourse than the other segment, the satellite. On the basis of a small corpus study, Matthiessen and Thompson (1988) report a strong (though not 1:1) relationship between nuclearity and hypotaxis, with most multinuclear relations being expressed using two main clauses, and most mononuclear relations using a main clause (nucleus) and a subordinate clause (satellite).

If main clauses are generally more central to, or prominent in, the discourse than subordinate clauses, it would explain why pronouns are preferentially interpreted as referring to the subject of an earlier main clause as compared to the subject of a more recent subordinate clause. However, considering that recency has also been shown to make content more accessible (Clark and Sengul, 1979; Cunnings et al., 2014; Streb et al., 2004), the pronoun interpretation rates found by Cooreman and Sanford (1996) seem very high. However, Miltsakaki (2011) suggests that complex sentences consisting of a main clause and a subordinate clause are treated as a single unit. If this is indeed the case, this would limit any 'boosting' effect of recency on the prominence of the subordinate clause. In a sentence-completion experiment, Miltsakaki (2003) compared the interpretation rates of ambiguous pronouns that were the subject of either a subordinate clause (3-a) or a main clause (3-b). She found that pronouns were more often used to refer to the preceding subject in the main-main (3-b) than in the main-sub condition (3-a).

(3) a. The beggar pushed the gentleman so that he ...
b. The boxer kicked the referee. As a result he ...

Miltsakaki (2011) argues that her results are consistent with the notion from Centering theory (Grosz et al., 1995) that topics are updated across sentences, and not within. In addition, they seem to be in line with sentence wrap-up effects during reading: a slowing down at the end of sentences, which has been proposed to at least in part be caused by readers integrating the information from the sentence with the preceding discourse (e.g., Just and Carpenter, 1980). It should be noted, however, that in Miltsakaki's experiment, syntactic configuration is confounded with one versus two sentences. It is therefore unclear what happens in contexts where main clauses do not appear as independent sentences (e.g., *The boxer kicked the referee and as a result he ...*), though if both syntactic configuration and sentence boundaries impact the prominence of preceding referents, the prediction would be that the pronoun resolution rate to the preceding subject should be somewhere between the observed rates for (3-a) and (3-b).

It thus appears that the specific configuration of a discourse mediates any recency effects on the accessibility of referents, but empirical evidence is fragmented and linguistic theories often focus on only one or two factors, leaving open the question of how different factors interact with each other. A general suggestion that arises from the literature is that the impact of subsequent linguistic material on the prominence of referents depends on the extent to which that linguistic material is seen as an integrated whole. In the literature on clause combining, factors that are seen as contributing to how integrated a clause is into its environment include not only syntactic subordination, but also whether a connective is present, and what kind of prosodic and/or typographic boundaries exist between the clauses (e.g., Gast and Diessel, 2012; Haiman and Thompson, 1988). Whether these factors all equally and in the same way contribute to a single dimension of integration is still an open question. If indeed several features of a discourse could impact the dimension of integration, it would mean that the relevant unit of measurement when talking about recency in discourse, even after taking into

account the Right Frontier, cannot straightforwardly be captured in number of clauses, sentences, discourse segments, etc.

In this study, we systematically investigate how the form of the discourse impacts the prominence of referents, and how different discourse features interact with each other in this respect. We start by replicating the effect of syntactic subordination, while teasing apart syntactic configuration and the presence of sentence boundaries (Experiment 1). However, in doing this, we introduce a potential confound between the presence of a sentence boundary and the presence of an explicit coherence marker. If the extent to which clauses are considered an integrated unit is a crucial factor (cf. Miltsakaki 2003), the presence of a connective could impact the prominence of referents as well, with explicitly marked coherence relations being treated as more integrated than implicit relations, which in turn would predict that intervening discourse segments without a connective lead to a stronger reduction in the prominence level of earlier referents than intervening discourse segments with a connective. We test this prediction in Experiment 2. Finally, in Experiments 3a and 3b, we switch to the spoken mode. Experiments 1 and 2, as well as the vast majority of related empirical work, examine written language. While some features are relevant to both the written and the spoken mode, others are not. The effect of syntactic subordination, however, can be expected to also be relevant to spoken language, but it is less obvious how the effect of sentence boundaries would transfer to the spoken mode. We therefore test whether the effects of syntactic subordination found in Experiments 1 and 2 can be replicated in spoken language, and we investigate the effect of the most obvious equivalent of typographic boundaries: high-level prosodic boundaries.

In all experiments, we use pronoun interpretation as a way of measuring the prominence of referents, where we expect higher resolution rates to more prominent referents (see von Heusinger and Schumacher 2019 for a discussion). Over the course of our experiments, we find that syntactic subordination, explicit coherence markers, and sentence boundaries all affect coreference patterns in a way that suggests that all three factors affect the extent to which two segments are treated as a single unit, and that they appear to do so in a straightforwardly cumulative way. While prosodic boundary tones have also been assumed to function as an integratedness cue (and are often seen as equivalents to typographic boundaries), the effects we find on coreference patterns are distinct from the other effects we examine. Between Experiment 3a and 3b, we replicate an unexpected interaction effect between syntax and prosody that suggests that the function of prosodic boundary tones in discourse should be reconsidered.

Since experimental materials largely overlap between all four experiments, Section 2 outlines the key features of our experimental items. Sections 3, 4, 5, and 6 respectively report on Experiment 1, Experiment 2, Experiment 3a and Experiment 3b. Each section contains a short discussion of the specific feature(s) under investigation and lays out detailed hypotheses for each experiment. The general discussion in Section 7 summarizes all results and reflects on what our study has contributed to our understanding of the relationship between recency and coreference, and, more generally, to our understanding of how the mental representation of a discourse is affected by the specific form of the discourse.

2. Experimental materials

2.1 Target items

In all experiments, the critical items comprised 24 short discourses. They always consisted of a biclausal construction followed by a continuation sentence with an ambiguous pronoun as subject

of a nonce verb predicate. Within the biclausal construction, each clause consisted of a different named human referent as subject, plus a predicate with an inanimate object (direct or prepositional). The subjects of the two clauses were of the same gender, so that the ambiguous pronoun could refer to either of the two subject referents.

The items were designed in such a way that independent of the conditions, the second clause was discourse-structurally subordinate to the first, so that both clauses would be accessible by virtue of being at the Right Frontier of the discourse. This is why the intended relation between the first two clauses was one of EXPLANATION: a backward causal relation at the content level (Frey, 2016; Sweetser, 1990). We took care to design the items in such a way that even when realized as two sentences with a full stop in between and without a connective, the meaning of the two clauses in sequence would make such a causal relation most likely. We aimed to avoid epistemic-level or speech act-level interpretations. This informed our choice of manipulations via connectives and prosody throughout the experiments. German has a number of discourse connectives that indicate a causal relation. We mainly focused on the prototypical subordinating connective *weil* with verb-final word order (all experiments) and the coordinating connective *denn* with verb-second word order (Experiments 1, 3). *Weil*-clauses with verb-final word order typically receive a content-level interpretation (Antomo and Steinbach 2010). We did not investigate *weil*-clauses with verb-second (main clause) word order. While they are frequent in spoken German, they carry prescriptive stigma in the written language (especially amongst introductory students of German) and are often used for epistemic- or speech act-level relations (Antomo and Steinbach 2010, Volodina 2011, Volodina 2014). For *denn*-clauses, some (e.g. Volodina 2014) have questioned whether they can encode content-level causal relations. Scheffler (2013), however, assumes that *weil* and *denn* share the same core meaning of causality, with the difference that this is conveyed as a conventional implicature in the sense of Potts (2004) in the case of *denn*, while it is asserted in the case of *weil*. We follow her and Wöllstein and Dudenredaktion (2022) in assuming that *denn* can indeed also be used for content-level causal relations, when this is contextually the most plausible reading. Regarding prosodic realization (Experiments 3a & 3b), it has been claimed that the degree of prosodic integration of causal clauses often correlates with their degree of syntactic integration and that less syntactic integration makes epistemic or speech act readings of the causal relation more available, while more integration favours a propositional reading (Pasch et al. 2003, Breindl 2009, Antomo and Steinbach 2010). Based on a corpus study, Volodina (2011, 2014) finds that there is a clear preference for a content-level reading with an integrated prosodic structure, while less integrated prosodic structures occur with all types of readings. As cues for prosodic integration she counts a rising boundary tone at the interclausal boundary, no or only a short pause there and a realization within a single intonation phrase. See the sections on Experiments 3a and 3b for how we operationalized (degrees of) prosodic integration. In sum, we assume that our item design in terms of content should make an epistemic or speech act reading of them less plausible. Item design was also intended to minimize any semantic bias for one or the other of the two subject referents. This is why we chose nonce verbs as the predicates in the continuation sentences, so that the meaning of those verbs would not influence the reference of the ambiguous pronoun. The experimental conditions differed in how the two clauses in the biclausal construction were linked syntactically (Experiment 1-3), typographically (Experiments 1-2), and prosodically (Experiment 3). Lists of all items for all experiments as well as the audio stimuli used in Experiments 3a and 3b are available in the accompanying online repository (`https://osf.io/r5jtk/?view_only=70308cb0c5554ed789359cf7f4759e8d`).

2.2 Filler items

In addition to the 24 target items, participants in each experiment were shown 16 filler items and 8 "catch" filler items. Fillers were similar to the experimental items in that the first two clauses also contained two subjects of the same gender, but their structure was freer, they featured a variety of different connectives, and contained both (discourse-structural) coordinating and subordinating discourse relations. Catch fillers were like the other filler items, except that the subjects of the first two clauses were of different genders so that the reference of the pronominal subject of the nonce verb in the continuation sentence was no longer ambiguous. Catch fillers were used to exclude data from inattentive participants.

3. Experiment 1

Experiment 1 investigates the effects of syntactic subordination / coordination and of type of typographic clause boundary on the prominence of referents, testing the assumption that both clause type (subordinate vs. main clause) and the perceived integration of clauses affect the prominence of discourse segments.

3.1 Conditions

Experiment 1 had three conditions for the critical items, see (4). They differed in whether the second clause (a) was separated by a comma, introduced by the causal connective *weil* and had verb-final word order [*weil*-condition], (b) was separated by a comma, introduced by the causal connective *denn* and had verb-second word order [*denn*-condition], or (c) was separated by a full stop and initially capitalized, not introduced by any connective, and had verb-second word order [2-sentence-condition]. The first clause was always the same in all three conditions. The first condition (a: comma, *weil* and verb-final word order in the second clause) is an instance of syntactic subordination, while the second (b: comma, *denn* + verb-second in the second clause) and third condition (c: full stop, no connective, verb-second) are both instances of syntactic coordination (cf. Scheffler, 2013). The two coordinated conditions (b-c) differed in the type of typographic clause boundary and, as a result, in whether the two clauses were presented as a single sentence or as two sentences, as well as in the presence vs. absence of an explicit causal connective.

(4) a. Nadja hat vegane Burger gekauft, **weil** Sabine kein Fleisch isst.
Nadja has vegan.PL burgers.AC bought because Sabine no meat.AC eats

b. Nadja hat vegane Burger gekauft, **denn** Sabine isst kein Fleisch.
Nadja has vegan.PL burgers.AC bought because Sabine eats no meat.AC

c. Nadja hat vegane Burger gekauft. **Ø** Sabine isst kein Fleisch.
Nadja has vegan.PL burgers.AC bought Sabine eats no meat.AC

Sie *daupte*.
She *dauped*.

'Nadja bought vegan burgers (because) Sabine does not eat meat. She *dauped*.'

Each experimental item was prepared three times, once in each condition, for a total of 3 x 24 = 72 items. These items were then distributed over three lists of 24 items via a Latin-square design.

Each list also contained the 16 filler and 8 catch filler items. Items in the lists were presented in randomized order and lists were distributed equally across participants. Participants were asked to identify the referent of the ambiguous pronoun (in (4) forced-choice between 'Nadja' and 'Sabine').

3.2 Predictions and hypotheses

If main clauses are more prominent than subordinate clauses (e.g., Bybee, 2001; Hooper and Thompson, 1973; Mann and Thompson, 1988; Matsuda, 1998; Smaby, 1974), and main-subordinate clause are treated as more integrated than main-main clause configurations (Miltsakaki, 2011), we expect the ambiguous pronoun to be more often resolved to the subject of the first clause (Ref1) in the *weil*-condition (4a) than in the two conditions with two main clauses (*denn*-condition, 4b; 2-sentence-condition, 4c). In addition, sentence boundaries have been suggested to form a stronger boundary between clauses than a clause boundary (Jasinskaja 2016; see also topic updating in Centering Theory (Grosz et al., 1995) and sentence wrap-up effects during reading (Just and Carpenter, 1980)). For German, Bredel (2008) even analyzes the full stop as an instruction to delete syntactic working memory and to start inter-sentential coherence operations, in contrast to the comma, which separates subordinate units but continues syntactic processing (though we suspect that a complete erasure of syntactic structure is maybe too strong a claim). If a full stop is indeed a stronger boundary than a comma, it would suggest that the two clauses in the *denn*-condition (4b) are seen as more conceptually integrated than the two clauses in the 2-sentence-condition (4c). The first clause in (4b) would then be less impacted by recency diminishing a clause's prominence level than the first clause in (4c), even though the surface distance is virtually the same. This would predict more resolutions to the subject of the first clause (Ref1) in (4b) than in (4c).

3.3 Participants

We recruited 106 German native speaking students at the University of Cologne as experimental participants. They participated for course credit and gave their informed written consent for us to anonymously use their experimental data for scientific purposes. For the statistical analysis, we excluded data from participants who gave at least one wrong answer to a catch filler item. This left data from 74 participants (65 female, 9 male; mean age 21.7 years, age range 18-46 years)[1] for the final analysis.

3.4 Procedure

The experiment was implemented online via PCIbex (Zehr and Schwarz, 2018). Participants accessed the experiment via a link and were randomly assigned to one of the three lists. At the beginning of the experiment, participants were informed about their rights and asked to give their consent for participation. They then received a short introduction to the experiment and could proceed to the experiment proper once they were ready. Items were presented in random order. For

1. A reviewer asks whether the low number of male participants after catch filler exclusion (here and in Experiment 3a) means that a disproportionally large number of male participants made mistakes with the catch fillers. That is not the case. There were 13 male participants in Experiment 1 before catch filler exclusion. In general, we recruited participants irrespective of gender and thus the gender imbalance is reflective of the specific populations the participants were drawn from. For Experiments 1 and 3a we recruited from students of the humanities at University of Cologne, who are more likely female than male, while for Experiments 2 and 3b we recruited from Prolific, whose German participant pool seems more equally balanced between female and male.

each item, participants were presented with the corresponding mini-discourse on their screen. Below the discourse on the screen, participants were asked for the referent of the pronominal subject of the nonce verb clause (e.g., *Wer daupte?* 'who dauped?'). They had to select between the two subject referents by clicking on one of two fields with their names (forced-choice). Since the reference of the pronominal subject was otherwise ambiguous, in the experimental items we expected our experimental conditions to influence participants' responses. In the "catch" filler items, on the other hand, pronominal reference was not ambiguous, so we expected participants to be able to respond according to grammatical gender without problems. The "catch" fillers were thus intended to detect participants that were not paying attention. Once participants made their choice, they were presented with the next item. They were not given a time limit within which they had to make their choice on each item, but the introduction asked them to make a quick selection based on their intuition.

3.5 Statistical analysis

Responses were coded as a binary choice to the first subject referent or not. A Bayesian generalized mixed-effects model with a binomial link function (to translate the binary choice) was fitted to the data using the package brms (Bürkner, 2017) in R (R Core Team, 2023), with weakly informative priors that do not skew the results in any direction but sufficiently reduce the parameter search space so that convergence issues do not occur (cf. McElreath, 2020; Vasishth et al., 2018).[2] A mixed-effects model is a statistical model that does not just take our fixed effects of interest into account but also factors that we cannot fully control, such as confounds arising from individual items or participants, by treating them as random effects. The model was treatment-coded, using the *weil*-condition as the intercept. The model included the three-level condition (*weil*/*denn*/2-sentence) as fixed effect, and random intercepts and slopes and their correlations for participant and item. In all experiments reported on here, we used Bayesian statistics because we can thus fit complex models with relative ease and because they provide us with a quantification of the uncertainty about our effects of interest via Credible Intervals (CrIs). In Bayesian statistics, the output of a mixed-effects model does not just consist of point values of the estimated parameters as in Frequentist statistics but of full probability distributions for each parameter estimated (including each individual random effect parameter), the so-called posterior distribution. The parameter estimates that the model summary gives us (and which we report on as Estimates in our tables) are the median values of the relevant distributions. The 95%-Credible Interval indicates the interval within which values for the estimated parameter drawn from the posterior distribution lie with 95% certainty, given the model and the data. When the CrI does not include zero, we take that as strong evidence that the effect in question is reliable, again given the model and the data. In addition, we sometimes conducted one-sided hypothesis checks based on the posterior distribution. The hypothesis checks indicate the probability that the effect is either negative or positive, depending on the sign of the estimated parameter, based on the proportion of the posterior distribution that is negative or positive. Note that while the implementation of hypothesis checks is a nod to the Frequentist language of tests

2. In particular, for priors of class 'b', 'sd', and 'Intercept', we specified them as 'normal(0,5)', i.e., as a normal distribution with a mean of 0 and a standard deviation of 5. Since we're in log space (to code the binary choice data via continuous numbers), in probability space this translates to a prior expectation that about 95% of slope estimates will fall within the range of [0.000045; 0.99995] and be normally distributed with the mean at 0.5. This means that any effect will need to overcome the bias encoded in the mean that things are at chance level. So these priors make explicit the conservative prior assumption that no effect is present, which is implicit to Frequentist thinking.

with significance thresholds, the logic behind it is different. A Bayesian hypothesis check simply compares aspects of the posterior distribution. For example, when we make a hypothesis check of whether the effect of parameter A is larger than that of parameter B, we compare the posterior distributions of the estimate of A to that of B. What we get is an indication of how likely it is on average that A is larger than B and how far apart on average a posterior sample of A is from one of B. This means that except for marginal cases, it also rarely makes sense to make a two-sided "test": since the estimates are measures of central tendencies of posterior distributions, their relative position (whether one is larger than the other) already indicates the overall relationship between them, and this is the direction that it makes sense to "test", i.e. to quantify how many of the posterior samples of the effect with the higher median are actually larger than those with the lower median. There are also implications for whether the equivalent of pairwise comparisons, which often means multiple testing in Frequentist statistics, requires an alpha-level adjustment: since all we do with Bayesian hypothesis tests is compare the posterior distributions of different effects, we do not actually test anything multiple times in the Frequentist sense of comparing a model-obtained point value estimate against a statistic, with the danger of a false positive increasing each time. All of these measured relationships between the different aspects of the posterior distribution are already there at the time the model is fully built, and thus they are all equally contingent on this model, of this data. This is why we use the qualification "given the model and the data" when talking about whether an effect is reliable. For the population as a whole, we can then take our results in the Bayesian spirit: as a piece of evidence for or against some hypothesis, whose value is determined by the strength and quality of the evidence and our prior domain knowledge. Data and R scripts for all experiments are available at the accompanying repository on OSF, at `https://osf.io/r5jtk/?view_only=7fc7381ff19c4c9e886e10550b874185`.

3.6 Results Experiment 1

The model results for the fixed effects are given in Table 1. Since they are on the logit scale, for ease of interpretation we use estimated marginal means (via the package *emmeans*: Lenth 2023) to provide averaged values (the median posterior predictive value) and credible regions (the highest posterior density regions, HPD) for the conditions on the response scale in Table 2 and Figure 1. In all our experiments reported on here, the median posterior predictive value can be understood as the average value of resolution to Ref 1 for the condition (combination) given. The HPD indicates the values where 95% of the posterior distribution lie when translated to the response scale. Model results and one-sided hypothesis checks based on them show that the differences between all experimental conditions are statistically reliable (posterior probability > 0.95, cf. Table 3).

Condition	**Estimate**	**Error**	**95%-Credible Interval**
Intercept (= *weil*)	0.56	0.25	[0.08; 1.06]
denn	-0.4	0.16	[-0.71; -0.09]
2-sentence	-2.28	0.27	[-2.83; -1.77]

Table 1: Fixed-effects results of the Bayesian mixed-effects model for Experiment 1. Values given are on the logit scale.

Condition	Median posterior predictive value	HPD region
weil	0.636	[0.517; 0.747]
denn	0.539	[0.417; 0.664]
2-sentence	0.154	[0.078; 0.232]

Table 2: Estimated marginal means with highest posterior density (HPD, 95%) regions for the three conditions in Experiment 1 based on the Bayesian model results. The Median posterior predictive value indicates the average (median) resolution per condition, the HPD region the values where 95% of the corresponding posterior distribution lie when translated to the response scale. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

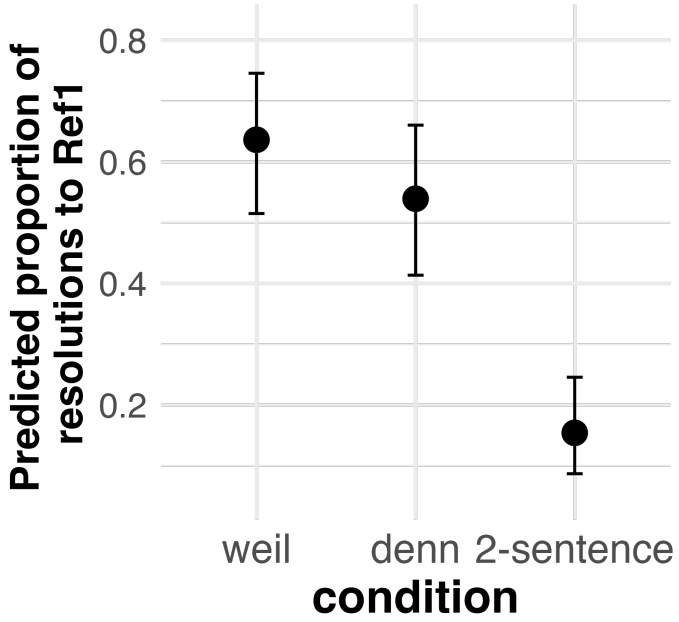

Figure 2: Estimated marginal means plus HPD regions (error bars) for the three conditions of Experiment 1 based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

Based on the model and the data, the *weil*-condition has the highest proportion of resolutions to the first-clause subject referent (mean posterior predictive value of 0.636; HPD = [0.517; 0.747]), followed by the *denn*-condition (mean posterior predictive value of 0.539; HPD = [0.417; 0.664]), and then the 2-sentence-condition (mean posterior predictive value of 0.154; HPD = [0.078; 0.232]). Since all differences between conditions are statistically reliable, the results confirm the predictions made for the experiment.

Hypothesis	Est.	Error	95%-Credible Interval	Evidence ratio	Posterior probability
weil > *denn*	-0.96	0.33	[-1.51; -0.43]	532.33	1
denn > *2-sentence*	-1.88	0.28	[-2.35; -1.44]	Inf	1
weil > *2-sentence*	-2.84	0.43	[-3.55; -2.13]	Inf	1

Table 3: One-sided hypothesis checks for the conditions in Experiment 1 based on the Bayesian model results. Estimate values given are on the logit scale.

3.7 Discussion

The statistical results confirm our predictions: resolution to the first-clause subject is higher in the syntactically subordinate *weil*-condition than in the two conditions with syntactic coordination. Within the two coordinate conditions, it is higher in the *denn*-condition with a single sentence than in the 2-sentence-condition. The results thus confirm that syntactic subordination affects prominence and that the perceived integration of clauses affects prominence, with typographic boundaries as a relevant signal for integration.

Looking at the results in more detail, we see that the difference within the two coordinated conditions is much larger than between the subordinate *weil*-condition and the coordinate *denn*-condition. It seems likely that this is (at least in part) due to the 2-sentence-condition being different from the *denn*-condition in more than a single factor: not only does it consist of two sentences separated by full stop rather than two clauses separated by comma, it also lacks the explicit causal connective that the *denn*-condition has. Explicit coherence markers can also impact the integration of clauses and sentences (see e.g., Koornneef & Sanders 2013 for a discussion). Experiment 2 tests whether the presence of explicit coherence markers impacts the prominence of referents independently of typographic boundaries.

4. Experiment 2

Experiment 2 aims to separate the role of 1) the presence/absence of an explicit coherence marker (connective) and 2) that of the sentence (as signaled by typography) as the relevant unit for recency in their effects on the prominence of subject referents.

4.1 Conditions

Experiment 2, like Experiment 1, had three conditions for the critical items. As in experiment 1, the first clause was the same in all conditions. In order to make the presence of an explicit causal connective independent from whether the clauses were combined by comma or full stop, we used the connective *nämlich* (≈ 'namely', 'you see'), which is causal and, because it is an adverb, compatible with both syntactically subordinating and coordinating constructions (Onea and Volodina 2011, Volodina 2014). Unlike *weil* or *denn*, *nämlich* cannot on its own occupy the initial position in a German clause, but it can cooccur with other connectives there or occur in sentence-medial position. In the first condition (a), [*weil-nämlich*-condition], the second clause was separated by a comma, introduced by the causal connective *weil* with verb-final word order and contained *nämlich*. In the second condition (b) [*nämlich*-condition], the second clause also contained *nämlich*, but was

separated by a full stop and had verb-second word order. The third condition (c) [2-sentence-condition] was the same as in the first experiment: the second clause was separated by a full stop, did not contain any connective, and had verb-second word order.

(5) a. Nadja hat vegane Burger gekauft, **weil** Sabine **nämlich** kein Fleisch isst.
Nadja has vegan.PL burgers.AC bought because Sabine nämlich no meat.AC eats
b. Nadja hat vegane Burger gekauft. Sabine isst **nämlich** kein Fleisch.
Nadja has vegan.PL burgers.AC bought Sabine eats nämlich no meat.AC
c. Nadja hat vegane Burger gekauft. **Ø** Sabine isst kein Fleisch.
Nadja has vegan.PL burgers.AC bought Sabine eats no meat.AC

Sie *daupte*.
She *dauped*.

'Nadja bought vegan burgers (because) Sabine does not eat meat. She *dauped*.'

Here again, the first condition is the only one where the second clause is syntactically subordinate to the first. In the second condition, the second clause is coordinate, and it is separated by a full stop, but the causal connection is still explicit because of *nämlich*, while in the third condition, it is implicit. The second condition here thus provides the intermediate step of combining a coordinate syntactic relation with an explicit connective and 2 sentences that the first experiment omitted.

4.2 Predictions and hypotheses

We again predict both syntactic subordination and orthographic boundaries to impact prominence, so that the *nämlich*-condition (5b) and the 2-sentence-condition (5c) should receive fewer resolutions to Ref1 than the *weil-nämlich*-condition (5a). We also hypothesize that explicit coherence markers have an effect on prominence, since they impact the integration of two clauses. Specifically, we expect resolution to Ref1 to be higher in the *nämlich*-condition (5b) than in the 2-sentence-condition (5c).

In comparison to Experiment 1, we expect the proportion of Ref1 resolutions in the *nämlich*-condition (5b) to lie between that of the the *denn*-condition (4b) and 2-sentence-condition (4c). Including *nämlich* in the first condition with *weil* (5a), in addition to making the interpretation of the relations as comparable across conditions as possible, allows us to assess the effect of *nämlich* alone, but we have no specific predictions for that.

4.3 Participants

We recruited 112 German speakers via Prolific as experimental participants. They were paid for their participation (1.50GBP for a task estimated to be 5-10 minutes; average completion time was 7.9 minutes) and gave their informed written consent for us to anonymously use their experimental data for scientific purposes. For the statistical analysis, we excluded data from participants who were not native speakers or gave at least one wrong answer to a catch filler item. This left data from 91 participants (45 female, 45 male, 1 other; mean age 29.7 years, range 20-50 years) for the final analysis.

4.4 Procedure

The experiment was conducted in the same way as Experiment 1.

4.5 Statistical analysis

Participants' response choices were again coded as a binary choice to first subject reference or not. We again fitted a Bayesian generalized mixed-effects model with a binomial link function to the resulting data, with weakly informative priors. The model was treatment-coded, using the *weil-nämlich*-condition as the intercept. The model included the three-level condition (*weil-nämlich*/*nämlich*/2-sentence) as fixed effect, and random intercepts and slopes and their correlations for participant and item.

4.6 Results

The model results for the fixed effects are given in Table 4. Estimated marginal means (using emmeans, Lenth 2023) for the conditions are given in Table 5 and Figure 3. Model results and one-sided hypothesis checks based on them show that again the differences between all experimental conditions are statistically reliable (posterior probability > 0.95, see Table 6).

Condition	**Estimate**	**Error**	**95%-Credible Interval**
Intercept (= *weil-nämlich*)	0.12	0.27	[-0.42; 0.66]
nämlich	-0.68	0.16	[-0.99; -0.37]
2-sentence	-1.31	0.19	[-1.69; -0.95]

Table 4: Fixed-effects results of the Bayesian mixed-effects model for Experiment 2. Values given are on the logit scale.

Condition	**Median posterior predictive value**	**HPD region**
weil-nämlich	0.529	[0.401; 0.662]
nämlich	0.362	[0.245; 0.493]
2-sentence	0.234	[0.147; 0.339]

Table 5: Estimated marginal means with highest posterior density (HPD, 95%) regions for the three conditions in Experiment 2 based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

Based on the model and the data, the *weil-nämlich*-condition has the highest proportion of resolutions to the first-clause subject referent (mean posterior predictive value of 0.529; HPD = [0.401; 0.662]), followed by the *nämlich*-condition (mean posterior predictive value of 0.362; HPD = [0.245; 0.493]), and then again the 2-sentence-condition (mean posterior predictive value of 0.234; HPD = [0.147; 0.339]). Since all differences between conditions are statistically reliable, the results confirm the predictions made for experiment 2.

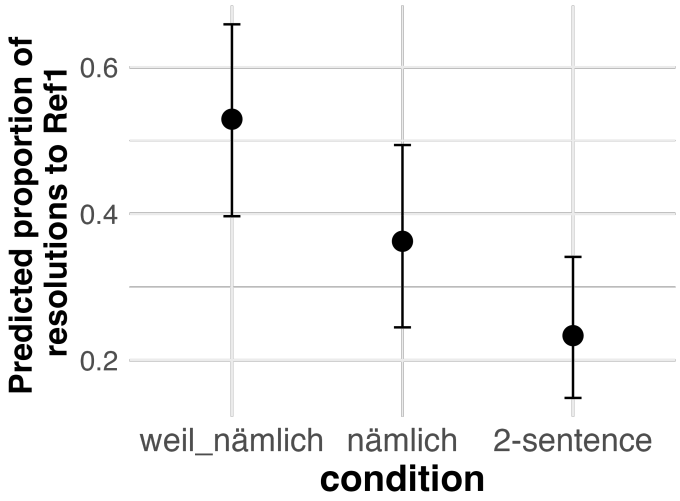

Figure 3: Estimated marginal means plus HPD regions (error bars) for the three conditions of Experiment 1 based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

Hypothesis	**Est.**	**Error**	**95%-Credible Interval**	**Evidence ratio**	**Posterior probability**
weil-nämlich > *nämlich*	-0.8	0.35	[-1.37; -0.24]	108.59	.99
nämlich > *2-sentence*	-0.63	0.19	[-0.94; -0.32]	1999	1
weil-nämlich > *2-sentence*	-1.42	0.38	[-2.04; -0.82]	1599	1

Table 6: One-sided hypothesis checks for the conditions in Experiment 2 based on the Bayesian model results. Estimate values given are on the logit scale.

4.7 Discussion

First of all, the results support those from Experiment 1: syntactic subordination and clause boundary type here also affect subject prominence, as seen in the difference between the *weil-nämlich*-condition and the *nämlich*-condition. Furthermore, our specific predictions for this experiment are also confirmed: the presence of explicit coherence markers has an effect on prominence that is independent of the effect of typographic sentence boundaries. We can see this from the fact that the second condition, with *nämlich*, has higher resolution to Ref1 than the third condition, which is identical except for the absence of *nämlich*.

In fact, the *nämlich*-condition lies about halfway (at 36% resolution to Ref1) between the *denn*-condition (54% resolution to Ref1) and the 2-sentence-condition (15% resolution to Ref1) of Experiment 1, which was predicted to be the case, since it differs from both by one manipulation factor (sentence boundary and explicit cueing of the relation). Of course, this can only be supporting evidence since we are comparing across experiments and the two 2-sentence-conditions in the two experiments do differ somewhat in their results (15% resolution to Ref1 in Experiment 1, 23%

resolution to Ref1 in Experiment 2). However, it clearly fits into the picture established by the more reliable comparisons between conditions within the experiments.

For the first condition, with *weil-nämlich*, we made no predictions beyond it having the highest resolution in Experiment 2, which was confirmed. Compared to the condition with only *weil* from Experiment 1, its mean resolution lies below that of that condition, which might be surprising given that there are two rather than one explicit markers of the causal relation in this condition. However, from the results in Table 5 we see that this condition also has the largest resolution range with an HPD interval of more than 0.25. We speculate that this might be because in its position directly following the subject, *nämlich* is ambiguous between an interpretation with wide scope in which its meaning relates to the whole clause and one with narrow scope in which it only relates to the subject. In the latter case, *nämlich* can have the effect that the subject is interpreted as a contrastive topic, usually coocurring with a pitch accent on the subject in spoken language (Karagjosova 2011). Our written materials however are ambiguous. Thus, the contrastive topic-reading would conceivably increase the prominence of the second-clause subject, while the wide-scope reading should behave similarly to the condition with only *weil*. Depending on how participants interpreted the written stimuli, they could therefore have assigned either comparatively low or high prominence to Ref1 in the *weil-nämlich*-condition.

The results from Experiment 1 and 2 were predicted by an account of clausal integration affecting prominence: The more the second clause is integrated into the first one forming a structural discourse unit, the more accessible the subject of the first clause is. In other words: the intervening second clause - if integrated - does not block the accessibility of the first subject. Both typographic boundaries and explicit coherence markers have been associated with clausal integration, and for both factors we found that the cue associated with a higher level of integration resulted in a higher resolution rate to the first-clause subject. We also observed these factors to operate independently of each other in a seemingly straightforward, cumulative way. In spoken language, there are no typographic boundaries, but there are prosodic boundaries, which are often seen as correlates of typographic boundaries and have also been suggested to function as integration cues. In the next experiments, we investigate how different prosodic boundary cues impact pronoun resolution in combination with syntactic subordination / coordination.

5. Experiment 3a

Experiment 3a aims to investigate the role that different prosodic boundaries play in interaction with syntactic cues on the prominence of referents, comparing three types of clause boundaries: no boundary, a high (rising) boundary, and a low (falling) boundary. We assume that the *no boundary*-condition represents a case where there is no or only a weak boundary between the two prosodic units mapping to the two clauses, while with both the *high boundary* and the *low boundary-conditions*, the increased pause length and the distinctive boundary pitch movement cue a stronger prosodic boundary. We are interested in the effect of prosodic boundaries because the degree of prosodic integration of a multiclausal complex is seen as an important cue for whether it is interpreted as (syntactically) more or less integrated, both crosslinguistically (e.g. Gast and Diessel 2012) and also specifically in the case of German causal clauses with *weil* (Pasch et al. 2003, Volodina 2011).

5.1 Conditions

Experiment 3a had a 2x3 design, crossing syntactic with prosodic factors, see (6). On the syntax side, it repeated the first two conditions from Experiment 1, with *weil* and *denn* as connectives, and verb-final and verb-second word order, respectively. On the prosody side, there were three conditions that manipulated both the type and the strength of the boundary between the two clauses: in the *low boundary*-condition (L%), pitch at the end of the first clause had a clear falling trajectory, followed by a pause of 250ms before the beginning of the second clause. In the *high boundary*-condition (H%), pitch at the end of the first clause had a clear rising trajectory, also followed by a pause of 250ms. In the *no boundary*-condition (Ø), pitch at the end of the clause did not follow a specific final trajectory and was rather suspended at medium height, followed by a shorter pause of only 30ms.

(6) a. Nadja hat vegane Burger gekauft {**Ø/H%/L%**} **weil** Sabine kein Fleisch isst.
Nadja has vegan.PL burgers.ACC bought because Sabine no meat.ACC eats

b. Nadja hat vegane Burger gekauft {**Ø/H%/L%**} **denn** Sabine isst kein Fleisch.
Nadja has vegan.PL burgers.ACC bought because Sabine eats no meat.ACC

Sie *daupte*.
She *dauped*.

'Nadja bought vegan.PL burgers.ACC because Sabine does not eat meat.ACC. She *dauped*.'

While Experiments 1 and 2 used written stimuli, Experiment 3a used audio stimuli. The items were recorded by one of the authors, who is trained in intonational phonetics and a German native speaker. The items were recorded in such a way that for each clause, there was only a single pitch accent that was always located on the object or, in one case, on the particle of a particle verb. The wording of the items was also slightly changed from those used in the first two experiments such that the first clause was always in the perfect tense, with the participle as the final word (see the accompanying repository at `https://osf.io/r5jtk/?view_only=70308cb0c5554ed789359cf7f4759e8d` for a full list of items and audios of all the experimental stimuli). This was done so that the pitch movement of the pitch accent on the word preceding the clause-final participle and that of the boundary movement at the end of the participle were always clearly separate. To create the different conditions, a recording of the first clause in the low boundary-condition was used as a basis. Recordings of the participle without a final pitch movement and with a rising final pitch movement were then spliced onto that recording without the participle to create the *no boundary* and *high boundary*-conditions, respectively. For the second clause, two versions were created, corresponding to the two syntactic conditions (with *weil* and *denn*). The first-clause- and second-clause-recordings were then spliced together with the appropriate pauses inserted between them. Example pitch tracks from the weil-condition are given in Figure 4. The nonce verb sentence with the ambiguous pronoun and the question asking after its referent

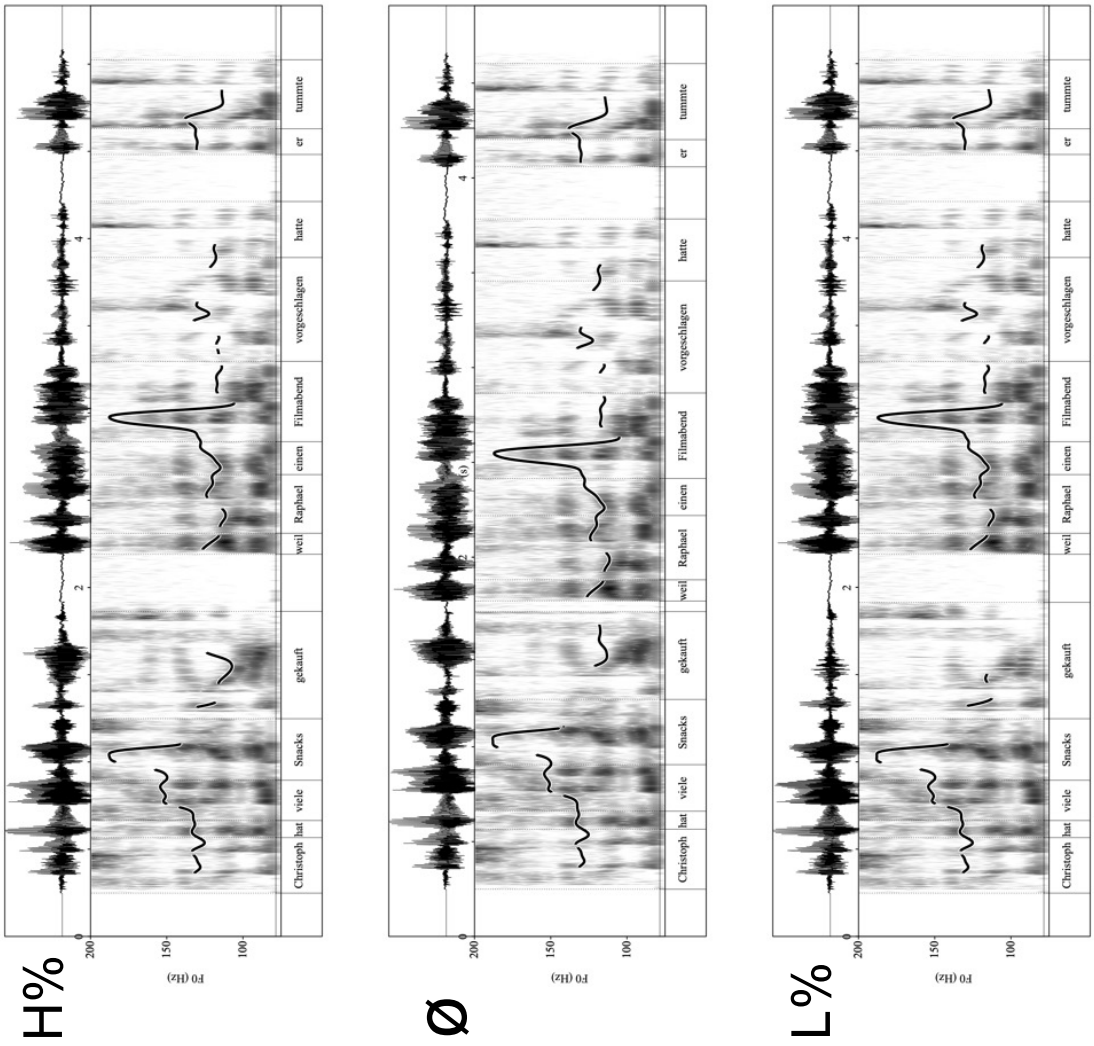

Figure 4: Waveforms, spectrograms and pitch tracks of experimental item *Christoph hat viele Snacks gekauft weil Raphael einen Filmabend vorgeschlagen hatte. Er tummte* 'Christoph bought many snacks because Raphael had suggested a movie night. He *tumm*ed.' (*weil*-condition), in the three prosodic conditions.

(*Wer daupte?* 'who dauped?') were also recorded. In each recording, following the full biclausal construction, first the nonce verb sentence was inserted after a pause of 250ms, followed by the question after another pause of 1000ms. All the recorded items were checked for naturalness by three German native speakers.

Apart from being audio rather than written, the differences in conditions and the slight alterations in the wording so that the participles were always final in the first clause, the experimental

items in Experiment 3 were the same as in Experiments 1 and 2. They were distributed across six lists via a Latin-square design and audio versions of the same fillers and catch fillers as in the first experiments were used.

5.2 Predictions and hypotheses

For the syntactic manipulations, the predictions were the same as in Experiment 1, i.e., more resolution to Ref1 in the *weil*-condition than in the *denn*-condition. For the prosodic manipulations, we had two hypotheses based on the literature:

1. **Boundary strength**: We hypothesized that in the *no boundary*-condition, having a weaker interclausal boundary than in both the *high*- and the *low boundary*-condition, the two clauses should be more integrated. This is based on the theoretical assumption of a prosodic hierarchy (Nespor and Vogel 2007), meaning that speech is ordered into prosodic units of increasing size, with corresponding boundary cues. While the boundaries of prosodic units clearly do not always coincide with syntactic boundaries, it is often assumed that there is at least a preferential correspondence, such that e.g. syntactic phrases match to phonological phrases, and syntactic clauses, larger syntactic units, match to intermediate or intonational phrases, larger prosodic units that are also marked by tonal movement at their boundaries (cf. Selkirk 2011). The assumption that stronger boundary cues (e.g. (longer) breaks, (more) final lengthening) correspond to stronger boundaries is encoded into many widely used prosodic transcription systems, including German ToBI (Grice et al. 2005), where it is also assumed that boundary tones belong to a prosodic unit that is at least an intermediate or intonational phrase. We therefore assume that both the shorter break and the absence of boundary tones mean that the boundary in the *no boundary*-condition is weaker. This should decrease the likelihood of the second clause being interpreted as independent (cf. Schubö et al., 2015). Specifically for *weil*-clauses, it has been claimed that if a biclausal complex with a postposed V-final *weil*-clause is realized with a "single intonation contour" rather than two separate ones, the weil-clause is more likely to be interpreted as syntactically integrated (Pasch et al. 2003). Both a shorter or non-existent break rather than a longer one and an intonational rise rather than a fall at the interclausal boundary are seen as cues for stronger prosodic integration (Pasch et al. 2003, Volodina 2011). We thus predicted that in terms of reference to Ref1, Ø $>$ H%, L%.

2. **Boundary type**: We further hypothesized that in the *high boundary*-condition the second clause should be interpreted as less independent than in the *low boundary*-condition, based on a tradition in the literature that has argued for meanings like "openness" or "non-finality" for high boundary tones and "closedness" or "finality" for low boundary tones (cf. e.g., Cruttenden, 1981; Grice and Baumann, 2007; Peters, 2018) and on the assumption that a rising interclausal boundary in such constructions indicates prosodic integration while a falling one indicates prosodic disintegration (Pasch et al. 2003, Volodina 2011). We thus predict that in terms of reference to Ref1, H% $>$ L%.

Overall, we thus predict in terms of reference to Ref1, *weil* $>$ *denn* and Ø $>$ H% $>$ L%. We did not predict an interaction between the syntactic and prosodic conditions.

5.3 Participants

We recruited 152 students at the University of Cologne as experimental participants. They participated for course credit and gave their informed written consent for us to anonymously use their experimental data for scientific purposes. For the statistical analysis, we excluded data from all participants who were not German native speakers, reported hearing problems, reported participating while in a noisy environment, reported having been disturbed more than once while participating, or gave more than one wrong answer to a catch filler item. This left data from 111 (99 female, 12 male; mean age 22.2 years, age range 17-48 years) participants for the final analysis. We decided on this slightly less strict catch filler criterion because 111 participants already leaves less than 20 participants per list on average.[3]

5.4 Procedure

Experiment 3a was conducted online via *Qualtrics*. Participants heard the question about the reference of the ambiguous pronoun as part of the recording for each item. Only then were they presented with the two choices corresponding to the two subject referents on their computer screen. Participants could listen to each recording only once and had to then complete the same task as in the previous experiments: choosing which referent the ambiguous pronoun in the nonce-word sentence referred to (forced-choice).

5.5 Statistical analysis

Participants' response choices were again coded as a binary choice to first subject reference. We again fitted a Bayesian generalized mixed-effects model with a logistic link function to the resulting data, with weakly informative priors. Since the experiment had a crossed design, the model was sum-coded, with *weil* and *denn* coded as 0.5 and -0.5, respectively, and *low boundary*, *high boundary*, and *no boundary* coded as [0.5; 0], [0; 0.5], and [-0.5; -0.5], respectively. It included the 2-level syntactic condition (*weil* vs. *denn*) and the 3-level prosodic condition (low boundary vs. no boundary vs. high boundary) as fixed effects as well as their interaction, and random effects for participants and items.

5.6 Results

The fixed-effects results of the model are given in Table 7. Table 8 gives the estimated marginal means (via emmeans) on the response scale for each level resulting from crossing the syntactic and prosodic conditions. The same data is visualized in Figure 5.

Based on the data and the model, the *weil*-condition has a higher proportion of resolutions to the first-clause subject referent than *denn* (estimate for *weil* $\rightarrow$ *denn* = -0.59, CrI = [-0.84; -0.34]). The model does not show any reliable differences between the prosodic conditions.

3. We also ran the statistical analysis with the dataset resulting from applying the stricter catch filler criterion of excluding all participants that got at least one catch filler wrong, which leaves only 84 participants. The results show the same tendencies as the ones presented in the main text.

Condition	Estimate	Error	95%-Credible Interval
Intercept (grand mean)	0.5	0.24	[0.03; 0.98]
syntax (weil → denn)	-0.59	0.12	[-0.84; -0.34]
prosody 1 (no boundary → low boundary)	-0.11	0.15	[-0.4; 0.2]
prosody 2 (no boundary → high boundary)	-0.10	0.16	[-0.4; 0.2]
syntax x prosody 1	0.34	0.3	[-0.24; 0.91]
syntax x prosody 2	-0.21	0.31	[-0.84; 0.37]

Table 7: Fixed-effects results of the Bayesian mixed-effects model for Experiment 3a. Values given are on the logit scale.

Syntactic cond.	Prosodic cond.	Median posterior predictive value	HPD region
weil	L%	0.658	[0.542; 0.782]
weil	Ø	0.717	[0.614; 0.815]
weil	H%	0.69	[0.569; 0.795]
denn	L%	0.559	[0.427; 0.685]
denn	Ø	0.57	[0.436; 0.693]
denn	H%	0.525	[0.394; 0.654]

Table 8: Estimated marginal means with highest posterior density (HPD, 95%) regions for the six levels of the crossed syntactic and prosodic conditions in Experiment 3a based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

5.7 Discussion

As can be seen from the model results and the main effect visualizations, the experiment only produced reliable evidence for an effect of the syntactic condition, but not the prosodic conditions. Therefore, none of the prosody-specific predictions made for this experiment can be confirmed.

The results for the six crossed levels (Table 8 and Figure 5), however, suggest that there is an unexpected interaction of the syntactic and prosodic conditions. This interaction seems to play out in two ways. Firstly, and most importantly, whereas the *high boundary*-condition seems to increase reference to Ref1 when combined with the syntactic *weil*-condition, compared to the low boundary condition, as predicted, it seems to instead slightly increase reference to Ref2 when combined with the *denn*-condition, in a quite unexpected fashion. Statistically, this interaction effect is not reliable according to the model, but the results nonetheless suggest that the prediction that H% should simply increase reference to Ref1 independent of the syntactic condition compared to L% is too simple.

Secondly, we can observe that the *no boundary*-condition seems to behave more like the *H%*-condition when combined with the *weil*-condition, but more like the *L%*-condition when combined with the *denn*-condition. This could suggest that it is not interpreted as a separate condition. Instead, we could speculate that in each syntactic context, one type of prosodic boundary is more natural and less marked than the other (i.e., H% in *weil* and L% in *denn*), and that the *no boundary*-condition is interpreted or perceived like this less marked boundary type in that syntactic context. There

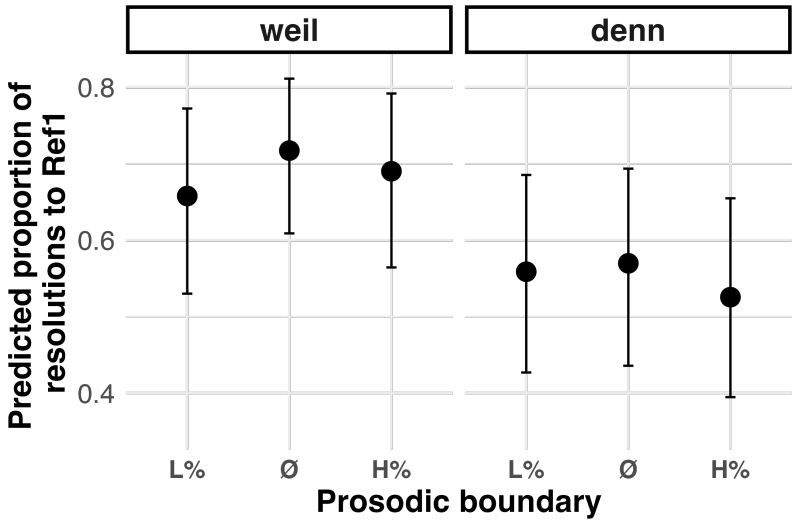

Figure 5: Estimated marginal means plus HPD regions (error bars) for the six levels of the crossed syntactic and prosodic conditions of Experiment 3a based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

is evidence that different types of causal clauses are indeed preferentially produced with different prosodic boundaries: Hu (2023) found that in English, boundaries preceding *because*-clauses were reliably produced with longer pause durations when the clauses were denoting subjective causality than when they denoted objective causality. For perception, Hu et al. (2023) found that in English clauses denoting forward causality with *so*, the prosody of the connective also influenced interpretation: when the connective was produced with longer duration and a more concave (rising) f0 contour, participants reliably chose a subjective continuation sentence compared to an objective one more often than when the connective was shorter and produced with a convex (falling) f0 contour. It remains an empirical question how applicable these results are for our German case at hand, even though the prototypical uses of *weil* and *denn* differ on the subjectivity dimension, with *weil* more often used in objective relations and *denn* in subjective relations (e.g., Pit, 2003). As already noted above, Volodina (2011) found a preference for prosodic integration which would often include a high boundary preceding the connective for *weil*-clauses with a content-level reading. For *denn*-clauses, on the other hand, Volodina (2014) states that they are preceded by a low boundary in the majority of cases. However, we will leave an exploration of this point for future research. Since the *no boundary*-condition conflates both boundary strength and boundary type, we decided to eliminate it in our next experiment in order to concentrate on the effect of boundary type alone and its interaction with syntax, prioritizing it in terms of statistical power.

An additional issue was that upon exploring our data, we found that individual items showed quite different behaviours with regards to the experimental conditions. Some items did not exhibit very much variation between the experimental conditions, in comparison to others, see Figure 6. Looking at those items exhibiting very little variation, we found that often, the semantics of the item made one of the two referents much more likely to be mentioned as subject of the continuation

sentence. For example, in item 8 (7), which showed very little variation and had a mean value of 0.87 (1 being 100% resolution to Ref1), it seems plausible that participants overwhelmingly chose referent 1 (Anna) as subject of the continuation sentence *sie zirfte* in Experiment 3a because that referent is both affected and present in the situation that is talked about, while referent 2 (Gabi) does not have to be present in the situation at all.[4]

(7) a. Anna hat kalte Spagetti gegessen weil Gabi die Mikrowelle
Anna has cold.PL spaghetti.ACC eaten because Gabi the.F microwave.ACC
kaputt gemacht hat. Sie zirfte.
broken made has she *zirfed*
'Anna ate cold spaghetti because Gabi broke the microwave. She *zirfed*.'

We found that about half of the items could conceivably contain a semantic bias in this sense. In order to have a clear separation criterion, for the next experiment we decided to replace all items meeting the following criterion: the interquartile range (IQR) of participant-averaged variation between conditions was less than 0.125 points on the response scale. This meant replacing 12, i.e., exactly half, of the items with items where both referents were physically present in the situation. We would like to stress that our aim was to reduce this semantic bias, which is an unwanted confound in our experiment. We created new items that would hopefully not be as noisy and thus allow us a less obscured look at whether there really is an effect of prosody. This implies that we expect any effect of prosody to be comparatively small and less robust against any strong semantic biases than the effect of syntax, which comes through despite of it.[5]

Overall, we were thus left with intriguingly unexpected but not fully reliable results for this experiment. With the next experiment (Experiment 3b), we intended to clarify these issues by gaining robust evidence for the possible interaction effect between syntax and prosodic boundary type.

6. Experiment 3b

Experiment 3b is a follow-up to Experiment 3a, again aimed at investigating the role that different prosodic boundaries play in interaction with syntactic cues on the prominence of subject referents. With the intention of obtaining more robust evidence for the possible interaction effect between the syntactic and prosodic conditions suggested by Experiment 3a, we made changes to both the conditions and the items. We omitted the *no boundary*-condition from Experiment 3a since it did not seem to behave independently from the other prosodic conditions and because that allowed us to focus on boundary type alone rather than boundary type as well as boundary strength. We also exchanged half of the items for ones that we hoped would allow for more variation in pronoun interpretation (see 5.7). We collected entirely new data from new participants for this experiment.

4. Though the contexts were different, Rohde et al. (2006) for example also show an effect of physical presence on coreference.

5. Indeed, similar differences in by-item variation can be found in Experiments 1 and 2. We have included plots in the supplementary data, including a plot of Experiment 1 with just the *weil* and *denn* conditions. These plots inform two observations: 1) Interpretation biases for each item are similar across experiments, underlining the assumption that the specific contents of an item plays an important role in resolving the ambiguous pronouns; 2) Even though the effect of *weil* vs. *denn* in Experiment 1 and 2 is fairly small and does not override the interpretation biases based on the semantics for each item, the effect appears to be consistent (non-noisy) enough to still show up in the main analysis (unlike the interaction effect in Experiment 3a).

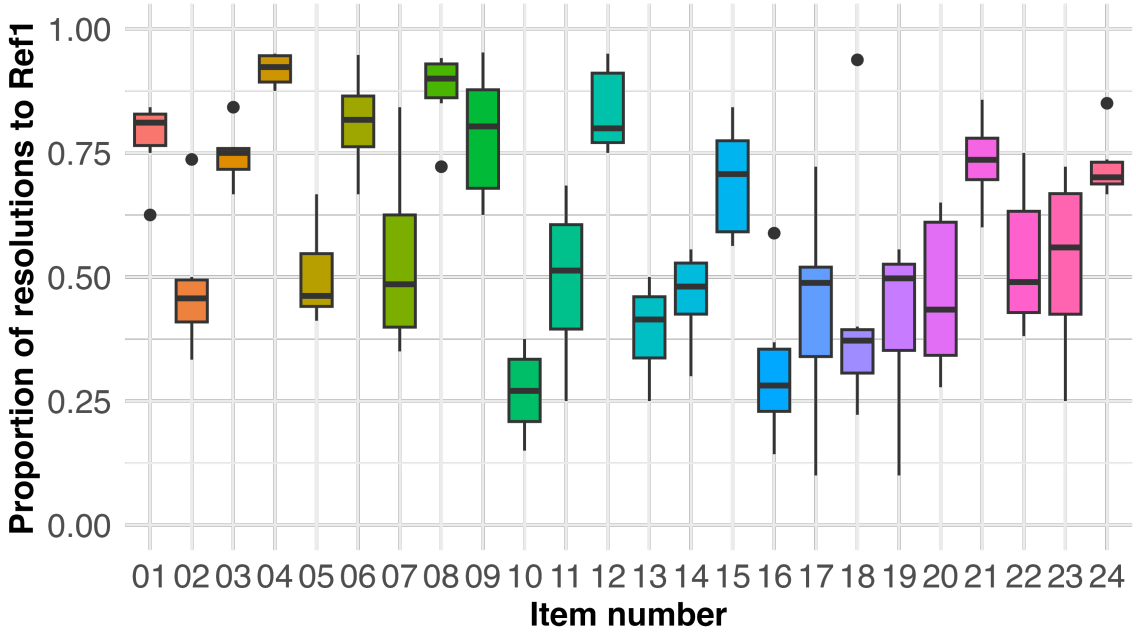

Figure 6: Boxplots of by-condition and by-participant variation for each experimental item in Experiment 3a.

In addition, we also recruited more participants per condition than before to gain robust results even when applying the strictest exclusion criterion of excluding all participants from the final analysis that got even one catch filler wrong.

6.1 Conditions

Experiment 3b omitted the prosodic *no boundary*-condition from Experiment 3a. Apart from that, the conditions stayed the same as in that experiment, thus creating a 2x2 experimental design.

6.2 Items

Compared to Experiment 3a, we replaced all items with an interquartile range (IQR) of participant-averaged variation due to the conditions of less than 0.125 points on the response scale, as described in the discussion section for Experiment 3a. This meant replacing 12 items, exactly half of the original ones. Since we had found many of them to plausibly bias interpretation towards one of the two subject referents via their semantics, we replaced them by ones that we hoped would be less biased semantically and more sensitive to our experimental manipulation.

Thus for example, we replaced (7) as item 8 from Experiment 3a with (8), where it is much more plausible that both referents are present in the situation that is currently talked about. We left both the referent names and the nonce verbs the same when replacing items. The new items were recorded and their recordings processed in the same way as those for Experiment 3a. Apart from these changes, the items were the same as in Experiment 3a.

(8) a. Anna hat den Ball angenommen weil Gabi nicht mehr
Anna has the.ACC ball.ACC received because Gabi not anymore
drangekommen ist. Sie zirfte.
it.reached is she *zirfed*
'Anna took the ball because Gabi couldn't reach it anymore. She *zirfed*.'

Items were distributed across four lists via a Latin-square design, together with the same fillers as in Experiment 3a.

6.3 Predictions and hypotheses

For the syntactic manipulations, the predictions were again the same as in Experiments 1 and 3a, i.e. more resolution to Ref1 for *weil* than for *denn*. For the prosodic manipulation, based on the tentative results from Experiment 3a and in contrast to the predictions from that experiment based on previous literature, we now predicted to find an interaction with the syntax for boundary type: namely that in the *weil*-condition, resolution to the first referent should increase in the *high boundary*-condition compared to the *low boundary*-condition, while in the *denn*-condition, resolution to the first referent should decrease in the *high boundary*-condition compared to the *low boundary*-condition.

6.4 Participants

We recruited 140 new German native speakers without hearing problems via the platform Prolific as experimental participants. They were paid for their participation (2.55GBP for a task that was estimated to take 13-18 minutes; average completion time was 14.8 minutes) and gave their informed written consent for us to anonymously use their experimental data for scientific purposes. For the statistical analysis, we excluded data from all participants who reported participating while in a noisy environment, reported having been disturbed more than once while participating, or gave at least one wrong answer to a catch filler item. This left data from 105 participants (49 female, 56 male; mean age 30.6 years, age range 20-50 years) for the final analysis.

6.5 Procedure

The experimental procedure was the same as in Experiment 3a.

6.6 Statistical analysis

We only used the data that was newly collected for this experiment in the analysis. As in the previous experiments, participants' response choices were coded as a binary choice to first subject reference and a Bayesian generalized mixed-effects model with a binomial link function was fitted to the resulting data, with weakly informative priors. Since we were expecting an interaction, the model was again sum-coded, with *weil* and *low boundary* each coded as -0.5, and *denn* and *high boundary* each coded as 0.5, respectively. It included the 2-level syntactic condition (*weil* vs. *denn*) and the 2-level prosodic condition (low boundary vs. high boundary) as fixed effects as well as their interaction, and random effects for participants and items.

6.7 Results

The fixed-effects results of the model are given in Table 9. Table 10 gives the estimated marginal means (via emmeans) on the response scale for each level resulting from crossing the syntactic and prosodic conditions. The same data is visualized in Figure 7.

Condition	Estimate	Error	95%-Credible Interval
Intercept (grand mean)	0.24	0.23	[-0.2; 0.71]
syntax (weil → denn)	-0.6	0.11	[-0.83; -0.38]
prosody (low → high)	0.03	0.14	[-0.24; 0.3]
syntax × prosody	-0.49	0.23	[-0.95; -0.04]

Table 9: Fixed-effects results of the Bayesian mixed-effects model for Experiment 3b. Values given are on the logit scale.

Syntactic cond.	Prosodic cond.	Median posterior predictive value	HPD region
weil	L%	0.6	[0.484; 0.709]
weil	H%	0.662	[0.541; 0.77]
denn	L%	0.512	[0.386; 0.633]
denn	H%	0.458	[0.345; 0.582]

Table 10: Estimated marginal means with highest posterior density (HPD, 95%) regions for the six levels of the crossed syntactic and prosodic conditions in Experiment 3b based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

Hypothesis	Est.	Error	95%-Credible Interval	Evidence ratio	Posterior probability
effect of high boundary >0 with syntactic condition = *weil*	0.27	0.17	[-0.01; 0.57]	17.39	.95
effect of high boundary <0 with syntactic condition = *denn*	-0.22	0.18	[-0.51; 0.08]	7.72	0.89

Table 11: One-sided hypothesis checks for the prosodic effect dependent on the level of the syntactic condition in Experiment 3b based on the Bayesian model results. Estimate values given are on the logit scale.

The results confirm and replicate the syntactic main effect already observed in Experiments 1 and 3a: the *weil*-condition has a higher proportion of resolutions to the first-clause subject referent than *denn* (estimate for *denn* = -0.36, CrI = [-0.66; -0.06]). The main effect due to the prosodic condition alone seems to be very small and not reliable. However, as the robust interaction effect in Table 9 and Figure 7 indicate, this seems to be because, as already hinted at by the results from

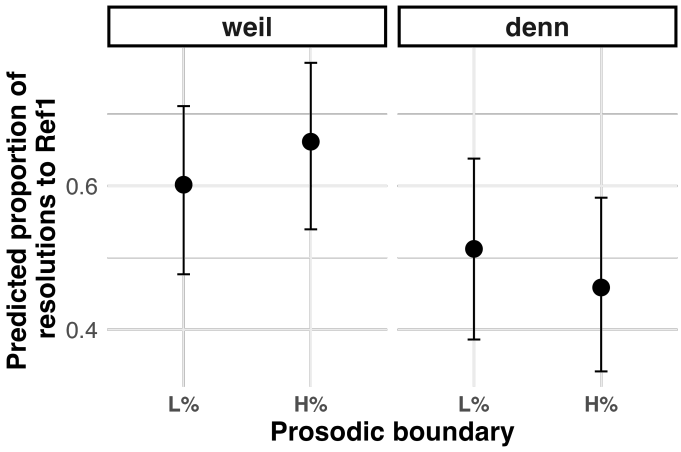

Figure 7: Estimated marginal means plus HPD regions (error bars) for the four levels of the crossed syntactic and prosodic conditions of Experiment 3b based on the Bayesian model results. Values given are on the response scale and reflect the probability of the first-clause subject referent being chosen.

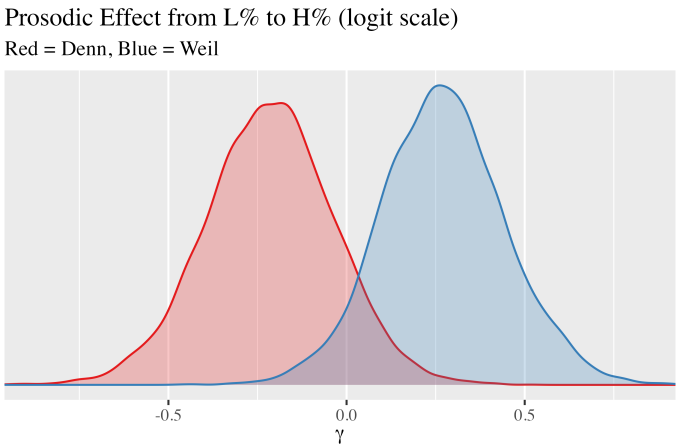

Figure 8: Posterior distributions of the conditional prosodic effect (from low to high boundary) in the syntactic *weil*-condition (in blue) and *denn*-condition (in red), based on the Bayesian model results of Experiment 3b. Values given are on the logit scale.

Experiment 3a and as predicted for this experiment, the effect due to the prosodic condition interacts with the syntactic condition. In order to corroborate this further, we ran a one-sided hypothesis check on the prosodic effect (moving from low boundary to high boundary) for each level of the syntactic condition, one for *weil* and one for *denn*. The results are given in Table 11 . As they indicate, the prosodic effect for *weil* is positive with a probability of 0.95, while that for *denn* is negative with a probability of 0.89. A look at Table 10 tells us that for *weil* this translates to an average increase in resolution to Ref1 from 0.6 to 0.66 from L% to H%. For *denn*, average resolution to Ref1 decreases from 0.512 to 0.458 from L% to H%. Additionally, we visualize the difference in the posterior distributions of the prosodic effect for *weil* and *denn* in Figure 7 . This allows us to see that while there is some overlap between the distributions, the majority of the distribution for the prosodic effect on *weil* is higher than that on *denn*. Accessing the posterior also allows us to calculate the number of individual posterior samples where, against the overall trend, the effect on *denn* is higher than that on *weil*. That is the case in only 1.6% of the posterior distributions. We consider this strong evidence that there is an interaction effect of prosody with syntax in our data with a trend for the prosodic effects to not only be of different strengths but to also go in opposite directions for the two levels of the syntactic condition. This also explains why there is no robust effect due to prosody alone.

6.8 Discussion

The main finding is that our prediction based on the preliminary results from Experiment 3a was confirmed in this experiment: we replicated the surprising interaction effect between prosody and syntax. Taken together, the results from Experiments 3a and 3b allow us to say with some confidence that prosodic manipulations at a relatively high level, viz. of phrasal boundary tones, affect the resolution of subject pronouns. Compared to the various syntactic manipulations, the effect is smaller, but crucially, it interacts with syntax in a surprising way. Whereas for a [main-sub]-configuration of two clauses connected by *weil*, a high interclausal boundary tone increases reference to the first clause subject referent compared to a low boundary tone, for a [main-main] configuration with *denn*, it decreases reference to the first clause subject. This seems to speak against equating interclausal boundaries with high tones with a higher degree of integration independent of context (cf. Schubö et al., 2015; Pasch et al., 2003; Volodina, 2011), if we assume that a less integrated second clause should be more independent and thus more reliably targeted as the most recently activated unit. Instead, the effect of the interclausal high boundary tone seems to interact with the upcoming structure: preceding a *weil*-clause, it strengthens the tendency of that clause to be perceived as dependent, resulting in fewer resolutions to its subject referent. This is in line with a straightforwardly cumulative effect of syntax and prosody. However, in the *denn*-condition, the effect of the high boundary tone is unexpected. Preceding a *denn*-clause, it appears to strengthen the tendency of that clause to be perceived as independent and results in more resolutions to its subject referent. In the following we briefly sketch a preliminary account that explains these findings in terms of a function of the high boundary tone as directing attention to what is coming up.

High boundary tones have long been recognized to convey rather abstract meanings variously labeled *openness*, *non-finality* or *incompleteness* (see e.g., Cruttenden, 1981; Gussenhoven, 2004). These labels all express the intuition that such high tones refer forward. It has recently also been found that boundary rises, compared to falls, increase attention orienting (involuntarily / subconsciously) during online processing (Lialiou et al., 2024a,b) and serial recall (Roehr et al., 2022)

for the larger domain with which they are associated. We speculate that this also means increased attention at what is coming next and how it is to be combined. Assuming this is the case, combined with the given that high boundary tones are probably more common before *weil* than before *denn* in everyday language use (Volodina, 2011) and thus more marked in the *denn*-condition, we suggest that when listeners encounter a *denn*-clause after the high boundary, they will be more attentive to the second clause. Since this is a coordinated and independent structure, listeners are more likely to represent it as separate from the main clause they just encountered before, compared to a situation where there was no high tone to increase their attention.[6]

This proposed account has the advantage of explaining our results and of making testable predictions about how prosody and syntactic structure should interact in other cases, e.g., with other connectives, or during online processing. We leave a more thorough elaboration for future work.

7. General discussion

Coreference seems to require antecedents to be at least somewhat recent. As the discourse continues, a mentioned referent's prominence seems to decrease, until it is no longer available to be picked up by a reduced referring expression. The relation between recency and prominence, however, does not appear to be straightforwardly linear, but is impacted by characteristics of the intervening discourse. In this study, we showed that an important feature of a discourse that can impact the prominence of referents is the extent to which intervening discourse segments are considered to be independent versus integrated with another discourse segment, and that the degree of integration of discourse segments is influenced by different factors. All experiments showed an effect of syntactic configuration, where an intervening main clause results in a higher drop in ambiguous pronouns resolved to an earlier referent than an intervening subordinate clause. In addition, Experiments 1 and 2 showed an effect of explicit coherence markers, where a discourse segment introduced by or containing a connective results in a higher proportion of ambiguous pronouns resolved to an earlier referent than a discourse segment that is simply juxtaposed with the preceding segment. Experiment 2 showed an effect of typographic boundaries, where an intervening clause presented as an independent sentence (by means of a full stop) results in a higher drop in ambiguous pronouns resolved to an earlier referent than an intervening clause that is separated from the preceding clause by means of a comma. Experiments 1 and 2 also demonstrated that these three factors appear to work in a straightforwardly cumulative way.

Experiment 3a and 3b investigated the effect of prosodic boundaries on coreference patterns, departing from the assumptions (based on the literature) that 1) the presence of a longer break and an intonational boundary tone are associated with a lower degree of integration between two clauses compared to a shorter break without a boundary tone, and that 2) a low boundary tone is associated with less integration between two clauses than a high boundary tone. The results from Experiment 3a, however, suggested that the effects of prosodic boundary tones on coreference were not as simple as these assumptions would predict. The *no boundary*-condition did not appear to behave as an independent condition, and instead behaved like the *high boundary*-condition in the *weil*-items and like the *low boundary*-condition in the *denn*-items. In addition, the prosodic boundary conditions appeared to behave differently depending on the syntactic condition (*weil* versus *denn*), though the

6. Although, as mentioned, high boundary tones are more common before *weil*, a similar mechanism could be at play here: the high boundary tone calls attention to the fact that the second clause should be integrated with the first clause, thus resulting in more prominence of the first clause compared to a sentence without such a boundary tone.

interaction was not statistically reliable. In Experiment 3b, we dropped the *no boundary*-condition and replaced the experimental items that showed little variation in the interpretation of the ambiguous pronoun (where presumably the meaning of the prompt shifted the interpretation bias of the item toward one of the two referents), to create more power to detect a possible interaction between prosody and syntax. The interaction between prosody and syntax suggested by Experiment 3a showed up statistically reliably in Experiment 3b: While a high boundary tone resulted in an increase in the proportion of pronouns resolved to the referent of the first clause in the *weil* (subordinate clause) condition, it led to a decrease in the *denn* (main clause) condition (this interaction showed up in addition to the main effect of syntactic configuration). These results confirm that even though high boundary tones have long been associated with a meaning of "openness" or "non-finality" and low boundary tones with "closedness" or "finality", they do not affect coreference patterns in the same way as other (non-prosodic) factors that seem to influence the integratedness of clauses (segments) in a discourse. This is especially interesting because prosodic boundaries, in particular intonationally marked boundaries, are often seen as correlates of typographic boundaries, in particular commas. That this is too simplistic a view for German has been shown by Kirchhoff (2016) (cf. Bredel, 2008). He proposes a model in which typographic commas in German are explained in part by intonational boundaries, and in part by syntactic conditions including one which places commas between subordinate clauses. However, since he adopts a syntax-prosody-mapping along the lines of Selkirk (2011) which essentially derives underlying intonational boundaries from syntactic ones, typographic boundaries are effectively based on syntax, with prosody only acting as an intermediary. Our results seem to suggest that at least in perception, intonation contributes cues to the construction of larger units of meaning in a way that is more independent from syntax than this view suggests. Since we found, against our initial predictions, that the choice of interclausal boundary tone interacts with the syntactic type of upcoming clause for coreference resolution, the information that the boundary tone provides is difficult to derive straightforwardly from syntax and seems to be truly additional. We have sketched an explanation for this interaction that makes recourse to assumptions about how (high) boundary tones direct attention to upcoming material during processing. However, it is speculative, and more research is needed both to corroborate or falsify this explanation, and to further investigate the possibly independent role of prosody in coreference resolution.

A reviewer raises a question about how unexpected the prosodic results are, and especially how unexpected their relatively weak effect size (about 5% resolution change in either direction) is, given the existing literature. If we consider the claims made about the importance of prosody for integration, not only is our observed interaction with syntax unexpected but also the small effect size. But we can also compare our findings with other studies on the effects of prosody on (pronominal) reference. There is a long-standing observation (going back to Akmajian and Jackendoff, 1970) that the difference between an unaccented and an accented pronoun can produce a preferred reference switch between two potential NP antecedents in English. Subsequent empirical research has confirmed this for English, but with qualifications: Itzhak (2013) (on subject pronouns), Taylor et al. (2013) and Mozuraitis and Heller (2017) (on object pronouns) observed a mean reference switch (with resolution changes from 20%-70% across different experiments) when the antecedent sentence and the anaphor sentence were in a PARALLEL coherence relation. For Spanish, Taylor et al. (2013), found a similarly sized effect of mean reference switch between using only a preverbal clitic pronoun and using a clitic pronoun doubled with a full postverbal pronoun, but prosodically accenting this full pronoun with a rising pitch accent did not have an additional effect. All three studies

also observed no comparable effect when the sentences were in a different coherence relation to each other. For (native) Italian, White et al. (2025) found an effect of about 25% resolution change that did not result in a mean reference switch between unaccented and accented conditions. For German, Hert (2023) did observe a reference switch of about 25% resolution change in a similar setup with the subject personal pronoun *er*, but no difference with the object personal pronoun *ihn*. Arguably however, our results should rather be compared with studies where not the pronoun itself was prosodically manipulated but the elements in the antecedent context. This was done in another study in Itzhak (2013) for English, where one of two NPs in the antecedent context was produced either with or without a rising pitch accent, followed by an ambiguous pronoun. She found an effect of about 13% resolution change. For German, Hert (2023) had a similar setup with two potential antecedent NPs followed by an ambiguous pronoun and a prosodic pitch manipulation on one of the NPs. In one experiment, she found a non-significant effect of about 5% and a significant effect of about 13% in another experiment for the personal pronoun *er*. For the demonstrative pronoun *der*, she found a stronger effect of around 25% resolution change. Given that in these studies, pitch (accent) manipulations on the antecedent NPs themselves only produced a relatively small effect, we think it is notable that our manipulation on only the boundary tones of the phrases in which the NPs are contained still produced an effect of comparable magnitude. This is interesting also in the light of findings (Grice et al., 2024) that pitch manipulations at the level of individual pitch accents vs. phrase-level boundary tones affect the serial recall of units differently, with high boundary tones improving the recall of the whole group they demarcate, while pitch accents only affect that of the individual unit they are realised on. Our results are compatible with one side of this domain-specific effect: (high) boundary tone manipulations affect the prominence of referents in the groups demarcated by them. But more research is clearly needed also to test its inverse: do pitch accent manipulations on individual referents affect the prominence of the larger grouping (e.g. a propositional discourse referent)? Overall, the varied experimental results summarized here show that (the size of) prosodic effects on pronoun resolution in contexts with two antecedents varies quite strongly, depending on the language, the type of pronoun, the discourse relation, whether pitch is manipulated on the pronoun itself or on the antecedent context, and potentially in the latter case, on the location of the pitch manipulation. More research is clearly needed to really understand how prosody interacts with other factors in pronoun resolution.

Our results show that when two referents are both accessible at the right frontier of discourse, the type of manipulations we have employed (syntactic sub-/coordination, presence/absence of connectives, typographic and prosodic boundaries) affect which one of them an ambiguous pronoun is preferentially resolved to. These manipulations all take effect at the level of the clause, and yet our results show that they affect the prominence of individual referents. Theoretically, the implicit assumption here is that the prominence of larger units (the clauses or rather their corresponding propositional discourse referents) affects the prominence of smaller units (the individual referents) contained within them. That is to say, that the smaller units "inherit" (some of) their prominence status from the larger units. We have empirically corroborated this intuitively plausible idea in a recent paper which also re-uses the data from Experiment 1 here (Buchholz et al., 2025). In the future, we intend to also conduct further empirical research on the inverse relation, i.e. how individual-level manipulations might affect clause-level referents, to provide this missing piece of the puzzle.

While our main focus in this paper was on the prominence of referents, our findings thus also extend to the question of how people build a discourse structure: which of the previous clauses do they attach a new clause to? This issue, while not trivial, has often been neglected within the field

of discourse coherence, much like other questions related to discourse segmentation (but see Hoek et al. 2021 for a rare empirical study on this topic). The current study proposes a general mechanism that guides the process of attaching segments to the preceding discourse: the more integrated with earlier segments the most recent discourse segment is, the more likely it is that a new segment attaches not to the most recent, but an earlier segment. Note that this mechanism is compatible with the Right Frontier Constraint, which posits that only those segments that subordinate the most recent discourse segment, and thus in a way discourse-structurally *contain* the most recent discourse segment, are available for easy attachment.

Our study was conducted in German, but the proposed mechanism, whereby the level of integration modulates the effect of recency on the prominence of clauses / discourse segments and the referents contained within those clauses, is assumed to generalize across languages. However, the exact contribution of specific features may differ between languages. German, for instance, has subordinate clause word order, which has been suggested to make the distinction between main clauses and subordinate clauses bigger compared to languages where word order between main and subordinate clauses is the same (e.g., English). The effect of grammatical subordination we found in the current study may therefore be bigger, or more consistent, than what we would find if we were to run the same study in English. In addition, in some languages, like for instance Japanese, connectives frequently appear in clause-final position, which may impact the effect of connectives on perceived integration found in the current paper. Finally, the current study only tested causal relations, but we again hypothesize the general mechanism to generalize across relation types. However, there may be characteristics of specific relation types or connectives that also impact the availability of clauses and the referents contained within them, see for instance (Buchholz et al., 2025) for a discussion of German *obwohl* 'although'-relations, in which the subordinate clause has been argued to express a proposition that is not asserted.

In sum, the current paper proposed a general mechanism of integration impacting the prominence of clauses and referents in a discourse, and identified several factors that impact the integration between clauses. Future work should further test the explanatory value of this account, and extend the research to other languages and other language features that may impact the extent to which clauses are considered an integrated unit.

8. Acknowledgements

This research was carried out as part of the CRC 1252 *Prominence in Language* at Cologne University (Project-ID 281511265), funded by the Deutsche Forschungsgemeinschaft (DFG). We gratefully acknowledge the support of the DFG. We also thank Stefan Baumann for consultation on designing the prosodic conditions, Max Hörl and Bodo Winter for help with statistics, and Nagihan Konuk, Robert Voigt, and Frederike Weeber for testing the acoustic stimuli for naturalness. Finally, we would like to thank three anonymous reviewers and the editor Amir Zeldes for very constructive feedback that helped us to substantially improve the paper. All errors remain our own.

References

Adrian Akmajian and Ray Jackendoff. Coreferentiality and Stress. *Linguistic Inquiry*, 1(1):124–126, 1970. ISSN 0024-3892. URL https://www.jstor.org/stable/4177532.

Mailin Antomo and Markus Steinbach. Desintegration und interpretation: Weil-v2-sätze an der schnittstelle zwischen syntax, semantik und pragmatik. *Zeitschrift für Sprachwissenschaft*, 29: 1–37, 2010. doi: 10.1515/ZFSW.2010.001.

Mira Ariel. *Accessing noun-phrase antecedents*. Routledge, 1990. doi: 10.4324/9781315857473.

Jennifer E. Arnold. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, 31(2):137–162, 2001. doi: 10.1207/S15326950DP3102_02.

Jennifer E. Arnold and Zenzi M. Griffin. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56(4):521–536, 2007.

Nicholas Asher. *Reference to abstract objects in discourse*. Number 50 in Studies in Linguistics and Philosophy. Kluwer Academic Publishers, 1993.

Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

Ursula Bredel. *Die Interpunktion des Deutschen*. Niemeyer, 2008.

Eva Breindl. Fehler mit System und Fehler im System - Topologische Varianten bei Konnektoren. In Marek Konopka and Bruno Strecker, editors, *Deutsche Grammatik - Regeln, Normen, Sprachgebrauch*, pages 274–308. De Gruyter, Berlin, 2009. doi: doi:10.1515/9783110217360.3.274.

Andreas Brocher and Klaus von Heusinger. A dual-process activation model: Processing definiteness and information status. *Glossa: A Journal of General Linguistics*, 3(1):1–34, 2018. doi: 10.5334/gjgl.457.

Timo Buchholz, Jet Hoek, and Klaus von Heusinger. Its prominence is her prominence: On the relationship between propositional and individual anaphoric reference. *Journal of Pragmatics*, 242: 36–59, 2025. doi: 10.1016/j.pragma.2025.03.013. URL `https://www.sciencedirect.com/science/article/pii/S0378216625000815`.

Joan Bybee. Main clauses are innovative, subordinate clauses are conservative: consequences for the nature of constructions. In Joan Bybee and Michael Noonan, editors, *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson*, pages 1–17. John Benjamins, 2001.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

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

Charles Clifton and Fernanda Ferreira. Discourse structure and anaphora: Some experimental results. In *Attention and performance XII*, pages 635–654. Routledge, 1987. doi: 10.4324/9781315630427.

Ann Cooreman and Tony Sanford. *Focus and syntactic subordination in discourse.* Research Paper no. RP-79, University of Edinburgh, HCRC, 1996.

H. Wind Cowles and Victor S. Ferreira. The influence of topic status on written and spoken sentence production. *Discourse Processes*, 49(1):1–28, 2012. doi: 10.1080/0163853X.2011.635989.

Alan Cruttenden. Falls and rises: meanings and universals. *Journal of Linguistics*, 17(1):77–91, 1981. doi: 10.1017/S0022226700006782.

Ian Cunnings, Clare Patterson, and Claudia Felser. Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language*, 71(1):39–56, 2014. doi: 10.1016/j.jml.2013.10.001.

Werner Frey. On some correlations between formal and interpretative properties of causal clauses. In Ingo Reich and Augustin Speyer, editors, *Co- and subordination in German and other languages*, volume 21, pages 153–179. Buske, 2016. doi: 10.1017/S1470542718000077.

Kumiko Fukumura and Roger P. G. van Gompel. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, 62(1):52–66, 2010. doi: 10.1016/j.jml.2009.09.001.

Kumiko Fukumura and Roger P. G. van Gompel. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, 26(10):1472–1504, 2011. doi: 10.1080/01690965.2010.506444.

Volker Gast and Holger Diessel. The typology of clause linkage: status quo, challenges, prospects. In Volker Gast and Holger Diessel, editors, *Clause linkage in cross-linguistic perspective. Data-driven approaches to cross-clausal syntax*, pages 1–36. de Gruyter Mouton, 2012.

Martine Grice and Stefan Baumann. An introduction to intonation – functions and models. In Jürgen Trouvain and Ulrike Gut, editors, *Non-Native Prosody: Phonetic Description and Teaching Practice*, pages 25–52. De Gruyter Mouton, 2007. ISBN 9783110198751. doi: 10.1515/9783110198751.1.25.

Martine Grice, Stefan Baumann, and Ralf Benzmüller. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, 2005.

Martine Grice, Michelina Savino, Petra B. Schumacher, Christine T. Röhr, and T. Mark Ellison. Rises on pitch accents and edge tones affect serial recall performance at item and domain levels. *Laboratory Phonology*, 15(1):1–26, 2024. doi: 10.16995/labphon.10473. URL `https://www.journal-labphon.org/article/id/10473/`.

Barbara J. Grosz, Aravind Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL `https://repository.upenn.edu/handle/20.500.14332/37528`.

Carlos Gussenhoven. *The phonology of tone and intonation*. Cambridge University Press, 2004.

John Haiman and Sandra A. Thompson. *Clause combining in grammar and discourse*, volume 18. John Benjamins Publishing, 1988.

Regina Hert. *The Role of Information Structure in Pronoun Resolution of Child, Adult L1 and Adult L2 Speakers*. PhD Thesis, University of Alberta, Alberta, 2023. URL https://era.library.ualberta.ca/items/1652100e-1719-4477-8d6e-d6160d648fad.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386, 2018.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted JM Sanders. Scolding the child who threw the scissors: Shaping discourse expectations by restricting referents. *Language, Cognition and Neuroscience*, 36(3):382–399, 2021.

Anke Holler and Lisa Irmen. Empirically assessing effects of the right frontier constraint. In António Branco, editor, *Anaphora: Analysis, algorithms and applications. DAARC 2007. Lecture Notes in Computer Science*, volume 4410, pages 15–27. Springer, 2007. doi: 10.1007/978-3-540-71412-5_2.

Joan B. Hooper and Sandra A. Thompson. On the applicability of root transformations. *Linguistic Inquiry*, 4(4):465–497, 1973.

Na Hu. *Speaking of causality. On the role of prosody in communicating subjective and objective causality in discourse*. PhD thesis, Utrecht University, 2023.

Na Hu, Aoju Chen, Hugo Quené, and Ted J. M. Sanders. The role of prosody in interpreting causality in English discourse. *PLOS ONE*, 18(6):1–22, 06 2023. doi: 10.1371/journal.pone.0286003.

Inbal Itzhak. *The influence of lexical biases and prosodic information during online interpretation of ambiguous pronouns*. PhD thesis, McGill University, Montreal, 2013. URL https://escholarship.mcgill.ca/concern/theses/dn39x471q. Publisher: McGill University.

Katja Jasinskaja. Not at issue any more. Manuscript, University of Cologne, 2016.

Marcel A. Just and Patricia A. Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980. doi: 10.1037/0033-295X.87.4.329.

Elena Karagjosova. Discourse particles, discourse relations and information structure: The case of nämlich. *International Review of Pragmatics*, 3(1):33–58, 2011. doi: 10.1163/187731011X561018.

Frank Kirchhoff. Interpunktion und intonation. In Ulrike Domahs and Beatrice Primus, editors, *Handbuch Laut, Gebärde, Buchstabe*, pages 398–417. De Gruyter, Berlin, 2016. doi: doi:10.1515/9783110295993-022.

Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023. URL https://CRAN.R-project.org/package=emmeans. R package version 1.8.6.

Maria Lialiou, Martine Grice, Christine T. Röhr, and Petra B. Schumacher. Auditory processing of intonational rises and falls in german: Rises are special in attention orienting. *Journal of Cognitive Neuroscience*, 36(6):1099–1122, 2024a. doi: 10.1162/jocn_a_02129.

Maria Lialiou, Jesse Harris, Martine Grice, and Petra B. Schumacher. Attention allocation to deviants with intonational rises and falls: Evidence from pupillometry. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46:667–674, 2024b. URL https://escholarship.org/uc/item/68q9c2r8.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

Kenjiro Matsuda. On the conservatism of embedded clauses. In Monika S. Schmid, Jennifer R. Austin, and Dieter Stein, editors, *Historical Linguistics 1997: Selected papers from the 13th International Conference on Historical Linguistics, Düsseldorf, 10–17 August 1997*, pages 255–268. John Benjamins, 1998.

Christian Matthiessen and Sandra A. Thompson. The structure of discourse and 'subordination'. In John Haiman and Sandra A. Thompson, editors, *Clause combining in grammar and discourse*, volume 18, pages 275–329. John Benjamins, 1988.

Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Routledge, 2020.

Eleni Miltsakaki. Effects of subordination on referential form and interpretation. *University of Pennsylvania Working Papers in Linguistics*, 9(1), 2003.

Eleni Miltsakaki. Not all subjects are born equal: A look at complex sentence structure. In Edward Gibson and Neal J. Perlmutter, editors, *The Processing and Acquisition of Reference*, pages 355–380. The MIT Press, 2011. ISBN 9780262295888. doi: 10.7551/mitpress/8957.003.0018.

Mindaugas Mozuraitis and Daphna Heller. Discourse coherence and the interpretation of accented pronouns. *Dialogue & Discourse*, 8(2):84–104, October 2017. ISSN 2152-9620. doi: 10.5087/dad.2017.204. URL https://journals.uic.edu/ojs/index.php/dad/article/view/10780.

Marina Nespor and Irene Vogel. *Prosodic Phonology*. De Gruyter Mouton, Berlin, Boston, 2007. ISBN 9783110977790. doi: doi:10.1515/9783110977790.

Edgar Onea and Anna Volodina. Between specification and explanation: About a german discourse particle. *International Review of Pragmatics*, 3(1):3–32, 2011. doi: 10.1163/187731011X561036.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Hermann Waßner. *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Number 9 in Schriften des Instituts für Deutsche Sprache. de Gruyter Mouton, 2003.

Jörg Peters. Phonological and semantic aspects of german intonation. *Linguistik Online*, 88(1), 2018. doi: 10.13092/lo.88.4191.

Mirna Pit. *How to express yourself with a causal connective: Subjectivity and causal connectives in Dutch, German and French*, volume 17. Brill, 2003.

Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5-6):601–638, 1988. doi: 10.1016/0378-2166(88)90050-1.

Christopher Potts. *The Logic of Conventional Implicatures*. Oxford University Press, 2004. doi: 10.1093/acprof:oso/9780199273829.001.0001.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/. Version 4.3.0.

Christine T. Roehr, Stefan Baumann, and Martine Grice. The influence of expectations on tonal cues to prominence. *Journal of Phonetics*, 94:101174, 2022. doi: 10.1016/j.wocn.2022.101174.

Hannah Rohde, Andrew Kehler, and Jeffrey L. Elman. Event structure and discourse coherence biases in pronoun interpretation. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 28, 2006. URL https://escholarship.org/uc/item/9js9w79w.

Tatjana Scheffler. *Two-dimensional semantics: Clausal adjuncts and complements*, volume 549. Walter de Gruyter, 2013.

Fabian Schubö, Anna Roth, Viviana Haase, and Caroline Féry. Experimental investigations on the prosodic realization of restrictive and appositive relative clauses in german. *Lingua*, 154:65–86, 2015. doi: 10.1016/j.lingua.2014.11.006.

Elizabeth Selkirk. The syntax-phonology interface. In John A. Goldsmith, Jason Riggle, and Alan C.L. Yu, editors, *The handbook of phonological theory*, pages 435–484. Blackwell, 2011. doi: doi:10.1002/9781444343069.ch14.

Richard M. Smaby. Subordinate clauses and asymmetry in English. *Journal of Linguistics*, 10(2): 235–269, 1974. doi: 10.1017/S0022226700006083.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548, 1994. doi: 10.1080/01690969408402130.

Judith Streb, Erwin Hennighausen, and Frank Rösler. Different anaphoric expressions are investigated by event-related brain potentials. *Journal of Psycholinguistic Research*, 33:175–201, 2004. doi: 10.1023/B:JOPR.0000027961.12577.d8.

Eve Sweetser. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge Studies in Linguistics. Cambridge University Press, 1990. doi: 10.1017/CBO9780511620904.

Ryan C. Taylor, Laurie A. Stowe, Gisela Redeker, and John C. J. Hoeks. Comprehension of Marked Pronouns in Spanish and English: Object Anaphors Cross-Linguistically. *Quarterly Journal of Experimental Psychology*, 66(10):2039–2059, October 2013. ISSN 1747-0218. doi: 10.1080/17470218.2013.773356. URL https://doi.org/10.1080/17470218.2013.773356.

Shravan Vasishth, Bruno Nicenboim, Mary E. Beckman, Fangfang Li, and Eun Jong Kong. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71:147–161, 2018. doi: 10.1016/j.wocn.2018.07.008.

Jorrig Vogels, Alfons Maes, and Emiel Krahmer. Choosing referring expressions in Belgian and Netherlandic Dutch: Effects of animacy. *Lingua*, 145:104–121, 2014. doi: 10.1016/j.lingua.2014.03.007.

Anna Volodina. *Konditionalität und Kausalität im Diskurs*. Narr, Tübingen, 2011. ISBN 978-3-8233-6559-4.

Anna Volodina. Kausale konnektoren. In Eva Breindl, Anna Volodina, and Ulrich Hermann Waßner, editors, *Handbuch der deutschen Konnektoren 2. Semantik der deutschen Satzverknüpfer*, pages 787–899. De Gruyter Mouton, Berlin, Boston, 2014.

Klaus von Heusinger and Petra B. Schumacher. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127, 2019. doi: 10.1016/j.pragma.2019.07.025.

Bonnie Webber. Discourse deixis: Reference to discourse segments. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122, 1988.

Lydia White, Heather Goad, Guilherme Duarte Garcia, Natália Brambatti Guzzo, Liz Smeets, and Jiajia Su. Pronoun interpretation in Italian: Exploring the effects of prosody. *Linguistic Approaches to Bilingualism*, 15(3):311–341, June 2025. ISSN 1879-9264, 1879-9272. doi: 10.1075/lab.22013.whi.

Angelika Wöllstein and Dudenredaktion. *Duden. Die Grammatik*. Dudenverlag, Berlin, 2022. ISBN 9783411914470.

Jeremy Zehr and Florian Schwarz. PennController for Internet Based Experiments (IBEX), 2018.