The effects of syntactic pressures and pragmatic considerations on predictive dependency formation

Maayan Keshev\textsuperscript{a,*} and Aya Meltzer-Asscher\textsuperscript{a,b}

\textsuperscript{a}Linguistics Department, Tel-Aviv University, Tel-Aviv, Israel; \textsuperscript{b}Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel

*corresponding author maayankeshev@mail.tau.ac.il
The effects of syntactic pressures and pragmatic considerations on predictive dependency formation

During sentence processing, comprehenders form predictions regarding the unfolding of the sentence. The current study is designed to tease apart the role of syntactic motivations, pragmatic considerations and contextual prediction in active dependency formation. Using both production and comprehension measures, we observe prediction for the resolution of a long-distance dependency, even in the absence of syntactic-licensing pressures. However, we find that predictive dependency formation is observed earlier when it is motivated by syntactic pressures. This suggests that the syntax plays a crucial role in this predictive process, even after controlling for pragmatic motivations. We propose that active dependency formation is faster or more binding in syntactically-motivated predictions (i.e. filler-gap dependencies) relative to pragmatically-motivated ones.

1. Introduction

During sentence processing, comprehenders form different predictions regarding the unfolding of the sentence. Yet, every act of predictive structure building or semantic composition requires resources, and puts the parser in greater risk of having to perform a costly reanalysis. Therefore, it might not always be cost-effective for the parser to engage in such strong predictive processes, and they may only come into play in cases where maintaining the structure in its incomplete form is very costly. The current study asks which types of pressures result in predictive dependency formation. We investigate the contribution of syntactic constraints, pragmatic motivations, and context to the prediction of long-distance dependencies.

Perhaps the most studied case of active dependency formation has been filler-gap dependencies. As elaborated on below, the parser predicts the resolution of filler-gap dependencies, and commits to the predicted analysis, to the degree that conflicting
subsequent input disrupts processing. In the current study, we test whether such prediction is unique to syntactically-motivated dependencies, or whether it takes place also when the discourse, rather than the syntactic context, assigns high probability to the occurrence of an expression co-referential with a previous antecedent.

1.1 Active formation of filler-gap dependencies

Filler-gap dependencies include a filler (italicized in 1) in a clause-initial position, and a gap (underlined) in the thematic position, where the filler is interpreted. Filler-gap dependencies provide an interesting example for the ability to project structure in advance of bottom-up evidence, since in incremental processing the position of the gap is not self-evident. Since the late 1980s it has been acknowledged that in these structures, the parser is eager to resolve the dependency, and postulates a gap prior to any direct evidence from the input. This has been attested over the years using different methods and in different languages.

One prominent consequence of this so-called "active filler" strategy (Frazier & Clifton, 1989) is the need for reanalysis when the parser's prediction for a gap is falsified (i.e. when another argument occupies the supposed gap position). This reanalysis effect, demonstrated by Stowe (1986, building on Crain & Fodor, 1985), was termed the Filled-Gap Effect (FGE). It is usually manifested in elevated reading times on the word occupying the potential gap position (the embedded object, *us*, in (1)) in sentences with filler-gap dependencies like (1a), versus baseline sentences with no dependency, e.g. (1b).

(1) a. My brother wanted to know who Ruth will bring us home to _ at Christmas.

b. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

Most studies of the "active filler" strategy consider it (implicitly or explicitly) a structurally-motivated principle. Some assume that this parsing strategy reflects a primary structural preference, which does not reduce to anything else (e.g. Frazier & Clifton's 1989
original formulation of the "active filler" strategy); others propose that it derives from a principle of minimizing structure (De Vincenzi, 1991; Inoue & Fodor, 1995); yet others argue that it reflects the urgent need of the parser to locate the canonical position of the filler and assign it a thematic role (Pritchett, 1992; Aoshima, et al., 2004), or results from a need to confirm that the dependency is well-formed (i.e. does not cross an island boundary, Wagers & Phillips, 2009).

However, non-syntactic accounts of the "active filler" strategy are also found in the literature. Altmann (1999) suggested that it results from the need to interpret (rather than to license) the sentence rapidly. Under this view, active gap-filling is a manifestation of the parser's efforts to extract as much informative content from the available partial input. Altmann suggests that anticipatory effects - parallel to predictively positing a gap - are present also when no filler-gap dependency is involved. The execution of active interpretation in other structures might suggest that the process is not dependent on specific grammatical pressures (see Kamide's 2008 review of various such experiments).

Lastly, another possible account for the "active filler" strategy can be formulated based on information structure premises. Describing the pragmatic discourse function of filler-gap dependencies, Kuno (1976, p. 420) postulated that "a relative clause must be a statement about its head noun". This suggests that the gap site in these dependencies must be a potential focus domain (i.e. in the part of the utterance which is asserted, rather than presupposed). Kuno's observation incited pragmatic analyses of different phenomena related to filler-gap dependencies, including island effects (Ambridge & Goldberg, 2008; Erteschik-Shir 1973, 1997; Goldberg, 2006), the distribution of resumptive pronouns (Ariel, 1999), and the availability of preposition stranding (Takami, 1988). It could be suggested that this pragmatic preference for filler-gap dependencies resolving in discourse-prominent positions can also account for the "active filler" strategy. The preference to resolve the dependency in a
discourse-prominent slot may lead the parser to predictively posit a gap as soon as possible, as early (matrix) positions are also more prominent.1

Crucially, the results from previous studies cannot decide between the different hypotheses. In previous experiments, the syntactic availability of a filler (differentiating the experimental and the baseline conditions) was confounded with availability of that referent. The studies used baseline structures that do not permit building a dependency at all, rather than eliminating only the syntactic motivation for creating a dependency (while still leaving semantic/pragmatic motivations intact). For example, in the baseline condition in Stowe (1986), exemplified in (1b) above, the wh-element is replaced with an if, thus removing a referent from the sentence. It is therefore impossible to know whether the slowdown on 'us' in (1a) relative to (1b) is because:

(i) The parser posited a gap in (1a) due to syntactic licensing pressures (lacking in 1b).

(ii) Attempting to attribute as much interpretation to the sentence fragment as possible, the parser assigned the theme role of 'bring' to 'who' in (1a), which is impossible in (1b) since this sentence lacks this additional NP which can receive the role.

(iii) Attempting to link the filler to a focus domain, the parser interpreted the filler as the direct object of the embedded verb, which is impossible to do in (1b) since the sentence does not contain a filler.

1 Another type of account for the "active filler" strategy suggests that it stems from resource limitations of the parser. These accounts present an architectural assumption, whereby the parser must maintain in working memory features of the unresolved dependency, which taxes processing resources (Wanner & Maratsos, 1978; Gibson, 2000). This pushes the parser to keep dependencies short, as a strategy for reducing processing load. However, if maintenance is difficult, why is the filler or other features of the dependency maintained active, more than other elements in the sentence, by default? Under the assumption that retrieval is available even without this active maintenance, the answer must resort to other motivations (such as the ones presented above) which require resolving the dependency.
To date, baseline sentences in experiments demonstrating the filled-gap effect either followed Stowe (1986) in lacking an antecedent altogether, or included structures in which the filler cannot be grammatically integrated as a complement of the verb, for example, due to subcategorization mismatch (see Table 1). Assuming that the parser does not predict ungrammatical dependencies, this type of manipulation similarly eliminates the possibility of reference to the antecedent at the relevant position. Therefore, these findings cannot exclude the contextual prediction account of the active gap-filling strategy.

<table>
<thead>
<tr>
<th>Study</th>
<th>Baseline condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stowe, 1986</td>
<td>If / whether clause</td>
</tr>
<tr>
<td>Boland et al., 1995</td>
<td>If / whether clause</td>
</tr>
<tr>
<td>FGE: John asked who Mary called us about __ yesterday evening. Baseline: John asked if Mary called us about her son yesterday evening.</td>
<td></td>
</tr>
<tr>
<td>Frazier &amp; Clifton, 1989 (Exp. 1)</td>
<td>Declarative sentence</td>
</tr>
<tr>
<td>Bourdages, 1992</td>
<td>Declarative sentence</td>
</tr>
<tr>
<td>Pickering, Barton, &amp; Shillcock, 1994</td>
<td>Declarative sentence</td>
</tr>
<tr>
<td>FGE: John knew who Mary called us about __ yesterday evening. Baseline: John knew that Mary called us about her son yesterday evening.</td>
<td></td>
</tr>
<tr>
<td>Clifton &amp; Frazier, 1989</td>
<td>Intransitive verbs</td>
</tr>
<tr>
<td>FGE: John knew what Mary should win at __. Baseline: John knew what Mary should excel at __.</td>
<td></td>
</tr>
<tr>
<td>Frazier &amp; Clifton, 1989 (Exp. 2)</td>
<td>Previous subject gap</td>
</tr>
<tr>
<td>FGE: John knew who Mary called us about __ yesterday evening. Baseline: John knew who __ called us about Mary yesterday evening.</td>
<td></td>
</tr>
</tbody>
</table>

2 Sussman and Sedivy (2003) targeted the question of contextual prediction in a visual-world eye tracking study. They monitored eye movements of participants scanning several object illustrations related to short background stories. They found predictive fixations on the picture corresponding to the direct object, at the offset of the verb in wh-questions (What did Jody squash the spider with?) relatively to yes/no questions (Did Jody squash the spider with her shoe?). However, as noted by the authors, fixations on this object in the wh-condition may also reflect utterance planning processes (Griffin & Bock, 2000) since in the wh-condition the object had to be named.
Aoshima et al., 2004  Case / category mismatch  
Lee, 2004  Case / category mismatch  
Wagers & Phillips, 2014  Case / category mismatch  
Wagers & Pendleton, 2016  Case / category mismatch  
Parker, 2017  Case / category mismatch  
FGE:  John met the student who Mary called us about __ yesterday.  
Baseline:  John met the student about whom Mary called us __ yesterday.  

Table 1. Baseline conditions used in studies which reported a filled-gap effect.

1.2 The rationale of the current study

The current study aims to tease apart the role of syntactic, pragmatic and contextual considerations in dependency prediction processes, and to uncover the motivation for active dependency formation in filler-gap dependencies. To test this, we investigate processing of the four Hebrew sentence types exemplified in (2).

(2)  
a. ha-menahel badak eyzo axot ha-metupal šixnea…  
The manager checked which nurse the patient persuaded…  
b. ha-menahel badak legaby ha-axot im ha-metupal šixnea…  
The manager checked regarding the nurse if the patient persuaded…  
c. ha-menahel badak im ha-axot im ha-metupal šixnea…  
The manager checked with the nurse if the patient persuaded…  
d. ha-menahel badak im ha-metupal šixnea³…  
The manager checked if the patient persuaded…  

The dependency in (2a) is of a filler-gap type, and is resolved before obtaining unequivocal evidence for a missing argument. On the opposite side of the scale, the sentence fragment in (2d) does not elicit a parallel prediction of a dependency at the point of the embedded verb, since it does not contain a parallel referent which the comprehender might expect to receive the theme role of persuaded. Sentences (2b-c) differ from (2d) in that their main clause includes an additional referent which can be used as a direct object for the verb persuaded. Therefore, in principle, the comprehender can predict that a pronoun co-referring

³ im ‘with’ and im ‘if’ are spelled differently in Hebrew (עם vs. אם, respectively).
with this antecedent will appear in that position (e.g. 'The manager checked regarding the nurse if the patient persuaded her…'). Crucially however, these sentences also diverge from (2a), as they do not involve a filler-gap dependency. Therefore, the antecedents in (2b-c) lack the syntactic motivation for active dependency formation.

There is also a critical distinction between sentences (2b) and (2c). These differ with regard to the information structure of the possible antecedent 'the nurse'. Sentence (2c) merely presents it as another discourse referent which might or might not be relevant later. However, (2b) presents it as a topic, a discourse-prominent antecedent within a regarding-phrase. Intuitively, Regarding-phrases in structures like (2b) require that the next clause will be a comment on the NP embedded in them. Since the clause is expected to be a statement about the antecedent in (2b), co-reference should be confined to prominent, matrix, positions. This is similar to what has been claimed about "as-for" topics (Kuno, 1976, 1987). According to Kuno (1976), this kind of discursive status is assigned to both "as-for" topics and fillers, and thus results in a similar preference for matrix co-reference in both cases, and in similar patterns of constraints. This is exemplified by the decreased acceptability of (3a-b) compared to (3c-d) (Falk, 2009).

(3)   a. As for TextMangler, I heard the rumor that Bill Gates uses it.

       b. Which word processor did you hear the rumor that Bill Gates uses?

       c. As for TextMangler, I heard that Bill Gates uses it.

       d. Which word processor did you hear that Bill Gates uses?

The different types of motivations for active dependency formation (syntactic requirements, information structure, or simple contextual availability), entail different predictions as to active formation of non-filler-gap dependencies:

(4)   a. **Syntactic licensing**: Active filler-gap dependency formation results from pressure to syntactically license the structure → No active postulation of co-referring elements
will take place when syntactic dependencies are not involved (i.e. predictive effects only in (2a)).

b. Discourse prominence: Active filler-gap dependency formation arises from preference for matrix positions as extraction sites, since the clause is a comment on the preceding topic (Kuno, 1976) → Prediction of a co-referring element will occur only in cases which involve referents (fillers or not) that require a comment referring back to them (i.e. predictive effects in (2a-b), but not in (2c)).

c. Maximal interpretation: Active filler-gap dependency formation is the result of the parser attempting to interpret as much of the sentence as possible (Altmann, 1999) → Active prediction of a co-referring element will occur for any contextually relevant referent (i.e. predictive effects in all three condition (2a-c)).

The following sections present three experiments that examine predictive dependency formation in the above cases, as reflected in acceptability, production, and online processing.

2. Experiments 1A & 1B: Acceptability

To make sure that the NP embedded in Hebrew regarding-phrases indeed constitutes a pragmatically prominent antecedent, which requires the next clause to be about it, we conducted two acceptability judgments experiments. In Experiment 1A we test whether a backward referring expression is required following regarding- or with- phrases (we compare the two dependency types to control for a possible general preference for repeated use of previous discourse referents. This experiment, therefore, tested the acceptability of sentences containing a regarding-phrase (as in (3b) above) and a with-phrase (as in (3c) above) with and without a subsequent pronoun. If a backward referring expression is pragmatically required following a regarding-phrase, the contrast between the pronoun and no-pronoun conditions should be greater in this structure.
In experiment 1B we test whether regarding-phares also require that the pronoun appears at an early/discourse prominent position. This experiment, therefore, tested the acceptability of sentences containing a regarding-phrase and a with-phrase with early (matrix) pronouns and late (embedded) ones. If there is preference for an early pronoun following a regarding-phrase, we predict the effect of the position of the pronoun to be greater in this structure than in the with-phrase structure.

2.1 Methods

2.1.1 Participants
Sixty-four native Hebrew speakers (according to self-report) volunteered for the study; 32 in Experiment 1A (mean age: 23.67, range 19-30), and 32 in Experiment 1B (mean age: 25.7, range 20-36). Fifty of the participants were monolingual and fourteen were bilingual speakers of Hebrew and either Russian, English or Spanish (distributed similarly between the two sub-experiments). Participants had at most one year of undergraduate linguistics education and were naïve to the purpose of the experiment.

2.1.2 Materials
Each of the two experiments included 20 experimental items of four conditions. The materials in Experiment 1A manipulated the type of antecedent (regarding- vs. with-phrase) and the realization of a pronoun referring back to it. The materials in Experiment 1B manipulated the type of antecedent and the position of the pronoun referring back to it (matrix/embedded). See Table 2 for sample items. Within each experiment, materials were divided into four lists in a Latin square design. Forty filler items of various acceptability statuses were added to each list (comprising a 60-item questionnaire).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sentence</th>
</tr>
</thead>
</table>

Regarding, Matrix pronoun (Exp 1A&B)

ha-šotrim badku legabey ha-soed im ha-xašuda
the-cops checked regarding the-customer whether the-suspect
mi-šxunat ha-mecuka šixne’a oto lidxf et
from.the-neighbourhood.of the-poverty convinced him to.push ACC
ha-melcarit ba-misada ha-yokratit
the-waitress in.the-restaurant the-luxurious

'The cops checked regarding the customer if the suspect from the poverty neighbourhood convinced him to push the waitress in the luxurious restaurant'.

Regarding, No pronoun (Exp 1A)

ha-šotrim badku legabey ha-soed im ha-xašuda
the-cops checked regarding the-customer whether the-suspect
mi-šxunat ha-mecuka šixne’a et ha-tabax
from.the-neighbourhood.of the-poverty convinced ACC the-cook
lidxf et ha-melcarit ba-misada ha-yokratit
to.push ACC the-waitress in.the-restaurant the-luxurious

'The cops checked regarding the customer if the suspect from the poverty neighbourhood convinced the cook to push the waitress in the luxurious restaurant'.

Regarding, Embedded pronoun (Exp 1B)

ha-šotrim badku legabey ha-soed im ha-xašuda
the-cops checked regarding the-customer whether the-suspect
mi-šxunat ha-mecuka šixne’a et ha-melcarit
from.the-neighbourhood.of the-poverty convinced ACC the-waitress
lidxf oto ba-misada ha-yokratit.
to.push him in.the-restaurant the-luxurious.

'The cops checked regarding the customer if the suspect from the poverty neighbourhood convinced the waitress to push him in the luxurious restaurant'.

With, Matrix pronoun (Exp 1A&B)

ha-šotrim badku im ha-soed im ha-xašuda
the-cops checked with the-customer whether the-suspect
mi-šxunat ha-mecuka šixne’a oto lidxf et
from the-neighbourhood of the poverty convinced him to push ACC ha-melcarit ba-misada ha-yokratit. the-waitress in the-restaurant the-luxurious.

'The cops checked with the customer if the suspect from the poverty neighbourhood convinced him to push the waitress in the luxurious restaurant'.

\[
\text{ha-šotrim badku im ha-soed im ha-xašuda}
\]

the-cops checked with the customer whether the suspect

\[
\text{mi-šxunat ha-mecuka šixne’u et ha-tabax}
\]

from the-neighbourhood of the poverty convinced ACC the cook

\[
\text{lidxof et ha-melcarit ba-misada ha-yokratit}
\]

to push ACC the-waitress in the-restaurant the-luxurious

'The cops checked with the customer if the suspect from the poverty neighbourhood convinced the cook to push the waitress in the luxurious restaurant'.

\[
\text{ha-šotrim badku im ha-soed im ha-xašuda}
\]

the-cops checked with the customer whether the suspect

\[
\text{mi-šxunat ha-mecuka šixne’u et ha-melcarit}
\]

from the-neighbourhood of the poverty convinced ACC the-waitress

\[
\text{lidxof oto ba-misada ha-yokratit.}
\]

to push him in the-restaurant the-luxurious.

'The cops checked with the customer if the suspect from the poverty neighbourhood convinced the waitress to push him in the luxurious restaurant'.

---

Table 2. Example stimuli from Experiments 1A and 1B.

2.1.3 Procedure

The questionnaires were built using Google forms. Participants filled the questionnaire online. They were instructed to rate the acceptability of the sentences on a 7-point scale, with
Order of the items was randomly generated for each participant.

2.1.4 Data Analysis
Statistical analyses were carried out on z-transformed data\(^4\) in R (R Development Core Team, 2015). A linear mixed-effects model was fitted separately for each experiment, with antecedent type, resolution type and their interaction as fixed effect predictors (factors were sum coded). In addition, two pairwise comparisons were conducted to test the effect of the resolution type, independently in for each antecedent type.

We conducted two types of analyses: a frequentist analysis (null hypothesis significance testing) and a Bayesian analysis using weakly informative priors. For the frequentist analysis, models were fitted using restricted maximum likelihood (REML) and p-values were extracted by Satterthwaite degrees of freedom approximation using the lmerTest package (Kuznetsova et al., 2017). We corrected the p-values of pairwise comparisons using the Bonferroni correction (for two comparisons). The Bayesian hierarchical models were fit using Stan (Carpenter et al., 2017), via the brms package (Bürkner, 2017).

Following Barr et al. (2013), a maximal random effect structure was constructed. Namely, in addition to random participant and item intercepts, we initially included correlated random slopes, by-participant and by-item, for each of the fixed effects as well as their interaction. However, due to convergence failure slopes were gradually removed from the frequentist analysis.\(^5\) The Bayesian models included a full random effects structure.

---

\(^4\) Within-subject z-transform is argued to eliminate variability that results from differences in scale biases across participants (Schütze & Sprouse, 2014).

\(^5\) Resulting in following models:

Experiment 1A: \(Z \sim \text{antecedent*resolution} + (1 + \text{antecedent*resolution} \mid \text{subject}) + (1 + \text{antecedent*resolution} + \text{antecedent} \mid \text{item})\)

Experiment 1B: \(Z \sim \text{antecedent*resolution} + (1 \mid \text{subject}) + (1 + \text{resolution} \mid \text{item})\)
We used regularizing, mildly informative priors (Gelman et al., 2014): a standard normal distribution $N(0, 1)$, was used as the prior distribution for the fixed effects and for the standard deviation parameters. For the intercept, we had a wider prior of $N(0,3)$. Finally, for the correlation matrices of the random effects, priors were defined using the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009). For each of the models, we sampled from the posterior distribution by running four MCMC chains at 2000 iterations each. The first half of the samples was discarded as warm-up samples. Convergence was checked using the R-hat convergence diagnostic and by visual inspection of the chains (Gelman et al., 2014). The code of both analyses and the associated data are available at https://osf.io/9XB2Y/ as supplementary materials.

2.2 Results

2.2.1 Experiment 1A

Z-transformed rating means by condition are presented in Table 3 and mean raw ratings are presented in Figure 1. Results of the frequentist and Bayesian analyses are summarized in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Matrix pronoun</th>
<th>No pronoun</th>
<th>Embedded pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1A:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regarding-phrase</td>
<td>0.36 (0.72)</td>
<td>-0.24 (0.79)</td>
<td>-</td>
</tr>
<tr>
<td>With-phrase</td>
<td>0.67 (0.55)</td>
<td>0.51 (0.73)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Experiment 1B:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regarding-phrase</td>
<td>0.23 (0.73)</td>
<td>-</td>
<td>-0.02 (0.82)</td>
</tr>
<tr>
<td>With-phrase</td>
<td>0.62 (0.55)</td>
<td>-</td>
<td>0.63 (0.65)</td>
</tr>
</tbody>
</table>

Table 3. Mean (SD) z-transformed acceptability ratings by condition, Experiments 1A-B.
The results of the frequentist analysis revealed a significant main effect of antecedent type \((t = -5.91, p < .001)\), such that dependencies with regarding-phrase antecedents were judged as less natural than those with with-phrase antecedents. We also observed a significant main effect of pronoun realization \((t = 5.56, p < .001)\), such that sentences with a pronoun were judged as more natural. Crucially, the interaction of the two factors was significant \((t = 2.77, p = .01)\). Pairwise comparisons (examining the effect of pronoun realization for each antecedent type separately) revealed a significant decrease in acceptability for the no-pronoun condition in regarding-type structures \((t = 5.19, p < .001)\). The corresponding comparison with-type revealed a marginal difference \((t = 1.98, p = .09)\). The Bayesian analysis produced similar results. The estimate and the credible interval (CrI) of the posterior distributions suggest a small interaction effect: the mean estimate of the posterior was 0.10, on the z-score scale, with CrI \([0.03, 0.18]\). In addition, the Bayesian model also exhibited evidence for a contrast between the two regarding-phrase conditions: the mean of the posterior was 0.60 with CrI \([0.36, 0.85]\).

<table>
<thead>
<tr>
<th></th>
<th>Bayesian</th>
<th>Frequentist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate [CrI]</td>
<td>Estimate</td>
</tr>
<tr>
<td>Main effect of antecedent</td>
<td>-0.27 [-0.38, -0.18]</td>
<td>-0.276</td>
</tr>
<tr>
<td>Main effect of resolution</td>
<td>0.20 [0.11, 0.29]</td>
<td>0.196</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.10 [0.03, 0.18]</td>
<td>0.106</td>
</tr>
<tr>
<td>Regarding: Pronoun - NP</td>
<td>0.60 [0.36, 0.84]</td>
<td>0.604</td>
</tr>
<tr>
<td>With: Pronoun - NP</td>
<td>0.18 [-0.05, 0.42]</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Table 4. Results of the frequentist and Bayesian analyses, Experiment 1A. For the frequentist analysis, we show the estimate, standard error, \(t\)-value and Bonferroni-corrected \(p\)-value of
the tested contrasts. For the Bayesian analysis we show the mean of the posterior distribution of each parameter of interest, together with a 95% credible interval (CrI).

Figure 1. Mean acceptability ratings (on a scale of 1-7) by condition, Experiments 1A-B. Error bar represent +/-1SE.

2.2.2 Experiment 1B

Z-transformed rating means by condition are presented in Table 3 and mean raw ratings are presented in Figure 1. Results of the frequentist and Bayesian analyses are summarized in Table 5. The results of the frequentist analysis revealed a significant main effect of antecedent type ($t = 5.54, p < .001$), such that dependencies with regarding-phrase antecedents were judged as less natural than those with with-phrases. Crucially, the interaction of the two factors was significant ($t = 2.55, p = .012$). Pairwise comparisons (examining the effect of dependency length for each antecedent type separately) revealed a significant difference between the short and long versions of regarding-phrase dependencies ($t = 2.58, p = .019$), such that sentences with earlier resolutions were rated higher. The corresponding comparison for the with-phrase dependency was non-significant ($t = -0.16, p > .99$). The Bayesian analysis produced similar results. The estimates and CrI of the posterior distributions suggest an interaction effect, which is small (mean = 0.07, on the z-score scale) but has a tight CrI [0.007, 0.13]. In addition, the Bayesian model also exhibited evidence for
some contrast between the two *regarding*-phrase conditions: the mean of the posterior was 0.26, on the z-score scale, with CrI [0.03, 0.50].

<table>
<thead>
<tr>
<th></th>
<th>Bayesian Estimate [CrI]</th>
<th>Frequentist Estimate</th>
<th>SE</th>
<th>t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effect of antecedent</td>
<td>-0.26 [-0.36, -0.17]</td>
<td>-0.264</td>
<td>0.025</td>
<td>-10.5</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Main effect of resolution</td>
<td>0.06 [-0.03, 0.16]</td>
<td>0.060</td>
<td>0.043</td>
<td>1.38</td>
<td>.18</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.07 [.007, 0.13]</td>
<td>0.069</td>
<td>0.025</td>
<td>2.75</td>
<td>.006</td>
</tr>
<tr>
<td><em>Regarding</em>: Matrix-Embedded</td>
<td>0.26 [0.03, 0.50]</td>
<td>0.259</td>
<td>0.100</td>
<td>2.58</td>
<td>.02</td>
</tr>
<tr>
<td><em>With</em>: Matrix-Embedded</td>
<td>-0.01 [-0.25, 0.22]</td>
<td>-0.018</td>
<td>0.100</td>
<td>-0.18</td>
<td>&gt; .99</td>
</tr>
</tbody>
</table>

Table 5. Results of the frequentist and Bayesian analyses, Experiment 1B. For the frequentist analysis, we show the estimate, standard error, t-value and Bonferroni-corrected p-value of the tested contrast. For the Bayesian analysis we show the mean of the posterior distribution of each parameter of interest, together with a 95% credible interval (CrI).

### 2.3 Discussion

In Experiment 1, we tested the effect of the presence of a dependency and its length on acceptability ratings in sentences containing *regarding*- vs. *with*-phrase antecedents. The results of Experiment 1A suggest that *regarding*-phrases require a backward referring pronoun (in the subsequent clause), beyond the general preference for pronouns over new discourse referents. Moreover, the significant interaction in Experiment 1B indicates that a preference for early (matrix) co-reference (over late, embedded co-reference) also depends on antecedent type. We interpret these findings as evidence for a pressure to form an early co-referential dependency following *regarding*-phrases, over and above any such general preference in neutral contexts with no discourse prominent antecedents. In the following experiments we investigate the online effects of this factor.
Incidentally, it might also be noted that, in both experiments, participants consistently rated the regarding-type dependencies lower than the with-phrase sentences. This might suggest that the usage of regarding-phrases is marked. This markedness could result from these structures' overall lower frequency in the language.\(^6\) In addition, there might be some additional discursive preference which applies to regarding-phrases sentences. It is possible that regarding-phrases are usually licensed by a fuller context than that available in one-sentence items such as those used in our experiments.

3. Experiment 2: Production

Although they provide a somewhat delayed measure of processing, sentence completion tasks are indicative of prediction as they allow manifestation of the predicted structure at the relevant point in the sentence. Experiment 2 thus aimed to examine how the predictability of different dependency types is reflected in production. The experiment included preambles of four conditions: wh-questions, assumed to present a syntactically active antecedent; regarding-phrase structures, assumed to present a discourse-prominent topic requiring a comment; with-phrase sentences, with an antecedent presented as a non-obligatory argument of the matrix clause; and bare if-questions, which eliminated the antecedent from the initial part of the sentence. We expected the production rate of co-referring elements to vary as a

---

\(^6\)A corpus survey indeed reveals that sentences of the form of our with-phrase interrogative constructions are much more common than the corresponding regarding-phrase ones. We tested the frequency of the structures in the HeTenTen corpus, a web-corpus of about \(^{10^9}\) Hebrew tokens (https://www.sketchengine.co.uk/hetenten-corpus/), which is part of the TenTen family of corpora of Sketch Engine (Jakubiček et al., 2013). We found 16,591 instances of sentences containing a verb immediately followed by a with-phrase, which in turn is followed by an interrogative within the subsequent four words. The corresponding search of regarding-construction yielded only 377 instances. A Similar ratio was revealed when limiting the search to the verbs used in our materials (with-construction: 1,214 instances; regarding-construction: 56 instances.
function of dependency prediction strength, such that antecedents which entail a dependency more strongly are expected to induce a higher rate of elements referring back to them.

3.1 Methods

3.1.1 Participants
Twenty-eight self-reported native Hebrew speakers, with a mean age of 30.31 (range: 20-53) volunteered for the study. Twenty-five of them were monolingual, and three were bilingual of Hebrew and either Arabic, Spanish or Italian. Participants were either speakers without any training or education in linguistics, or undergraduate linguistics students who were naïve to purpose of the experiment.

3.1.2 Materials
Twenty sets of sentence-initial fragments in Hebrew were created for this experiment, each containing four conditions, manipulating the type of antecedent provided in the sentence fragment (see Table 6 for examples). Sentence fragments presented either a wh-question (using a which-NP phrase); an if-question preceded by a regarding-phrase; an if-question preceded by a with-phrase which was a non-obligatory argument of the matrix verb, providing it a role of addressee to which the embedded question is directed; and bare if-questions, which eliminated the antecedent from the initial part of the sentence, leaving only one referent (the subject) in the matrix clause, and one in the embedded question. In all four conditions, the sentence was truncated following the verb of the embedded question.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which</td>
<td>ha-mora  badka eyze talmid ha-biryonim me-ha-kita</td>
</tr>
<tr>
<td></td>
<td>the-teacher checked which student the-bullies from-the-class</td>
</tr>
</tbody>
</table>

7 This was mostly an indirect object headed by the Hebrew counterpart of the preposition with, but in several sentences, it was realized as the direct object of the matrix verb.
Table 6. Example stimuli for Experiment 2.

3.1.3 Procedure

The questionnaire was built using Google forms, and participants filled it online. They were instructed to complete the sentence fragments as they wish, and to make sure they add at least one word.

3.1.4 Coding

Sentence completions were excluded from the analysis if they gave rise to an ungrammatical sentence, suggesting that the participant misread some part of the sentence fragment (removing four productions, comprising 0.71% of the data). The rest of the sentences were categorized based on the referent used as the first object of the embedded verb. We examined
the rate at which participants produced, at that position, a backward referring element (BRE), either a gap or a pronoun, matching the critical antecedent.

A pronoun completion was considered a BRE if and only if the pronoun was the first argument produced for the provided verb and agreed in number and gender with the critical antecedent (7a). Pronouns which did not agree with the antecedent (her in 7b) or were not an argument of the critical verb (i.e. possessive pronouns modifying the verb's object, like his in 7c) were coded as "other pronouns". Productions which provided a lexical NP as the first argument of the verb were coded as NPs, even if a pronoun or a gap occurred as an additional argument later in the production (7d).

(7) The kindergarten teacher forgot regarding the boy whether the bullies asked
   a. … him to push the girl.
   b. … her to punish him.
   c. … his sister to push him.
   d. … the girls to ignore him.

Productions were considered as containing possible gaps only in the wh-question condition and were coded as BRE if and only if there was a missing argument position for the provided verb adjacent to it, and no other possibly missing argument position in the production (8a). Namely, if the production included another verb which could be transitive, the gap was assumed to be an object of that verb (8b). In this kind of case, the provided verb does not have an argument which can be coded as any of the previously presented categories. Such completions, as well as cases of missing arguments in conditions other than the wh-

---

8 This category was defined to allow comparison with the if-baseline, where the critical antecedent was missing. We included the NPs with possessive pronouns in this category rather than in the NP, in order to allow a conservative estimate of the contrast with the baseline condition (i.e. to assume the highest possible rate of such pronouns in that condition).
question condition, were coded as intransitive uses of the provided verb (we use here the term intransitive to indicate that the verb had only a sentential complement with no direct object).

(8) The kindergarten teacher checked which boy the bullies asked
    a. … to climb the fence.
    b. … to help.

Note also, that in the if-question condition, production of a BRE is impossible since no relevant antecedent is available in the input. Thus, we consider the production rate of a pronoun of any kind in this condition as a baseline to which to compare the production rate of BREs in the other conditions.

3.1.5 Data analysis

A logistic mixed-effects model was fitted in R, to the production data based on BRE completion (or pronoun completion in the if-question condition). We used a sliding contrast scheme to detect an increase in BRE completion rate for every condition relative to the previous level. We therefore report three pairwise comparisons: with-phrase condition vs. if-baseline, regarding-phrase condition vs. with-condition, and wh-question vs. regarding-phrase condition. We conducted two types of analyses: a frequentist analysis (null hypothesis significance testing) and a Bayesian analysis using weakly informative priors. For the frequentist analysis, p-values were extracted by Satterthwaite degrees of freedom approximation using the lmerTest package (Kuznetsova et al., 2017), and corrected using the Bonferroni correction (for three comparisons). The Bayesian hierarchical models were fit using Stan (Carpenter et al., 2017), via the brms package (Bürkner, 2017).

Following Barr et al. (2013), a maximal random effect structure was initially constructed (assuming random participant and item intercepts, we initially included random by-participant and by-item slopes and correlations). However, due to convergence failure
slopes were gradually removed from the frequentist analysis. The Bayesian models included a full random effects structure.

We used regularizing, mildly informative priors (Gelman et al., 2014): a standard normal distribution N(0, 1), was used as the prior distribution for the fixed effects and for the standard deviation parameters. For the intercept, we had a wider prior of N(0,3). Finally, for the correlation matrices of the random effects, priors were defined using the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009). For each of the models, we sampled from the posterior distribution by running four MCMC chains at 2000 iterations each. The first half of the samples was discarded as warm-up samples. Convergence was checked using the R-hat convergence diagnostic and by visual inspection of the chains (Gelman et al., 2014). Estimates and credible intervals were back-transformed (to percentage/rate of BRE production) for convenient interpretation of the effects. The code of both analyses and the associated data are available at https://osf.io/9XB2Y/ as supplementary martials.

### 3.2 Results

Distribution of BRE productions by condition is presented in Table 7 and Figure 2. We report pairwise comparisons between every two consecutive conditions, in order to test the contribution of each type of dependency to the rate of BRE completions. Results of the frequentist and Bayesian modelling are summarized in Table 8.

<table>
<thead>
<tr>
<th></th>
<th>BRE</th>
<th>Other pronouns</th>
<th>NP</th>
<th>Intransitive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wh-question</td>
<td>130</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>137</td>
</tr>
<tr>
<td>Regarding-phrase</td>
<td>99</td>
<td>13</td>
<td>20</td>
<td>7</td>
<td>139</td>
</tr>
<tr>
<td>With-phrase</td>
<td>99</td>
<td>8</td>
<td>29</td>
<td>4</td>
<td>140</td>
</tr>
</tbody>
</table>

9 Resulting in the following model:

\[
\text{BRE} \sim \text{antecedent} + (1|\text{subject}) + (1+\text{regarding}|\text{item})
\]
BREs were produced in the *with*-phrase condition at a significantly higher rate than that of pronoun completions in the *if*-question condition ($z = 5.13, p < .001$). We failed to detect a significant difference between the rate of BRE productions in *regarding*-phrase and *with*-phrase conditions. However, in wh-questions, BRE completions were significantly more common than in *regarding*-phrase sentences.

The Bayesian analysis produced similar results. The mean and CrI of the posterior distribution suggest an increase of 37% [21, 52] in *with*-phrase structures relative to the *if*-baseline. We did not observe a clear difference between *regarding*-phrase and *with*-phrase conditions: the mean of the posterior was 5% with CrI [-9, 19]. Lastly, estimates of the posterior distribution suggested an increase in BRE production in wh-questions relative to regarding-phrase sentences: the mean of the posterior was 34% with CrI [17, 50].

<table>
<thead>
<tr>
<th></th>
<th>Bayesian Estimate [CrI]</th>
<th>Frequentist Estimate</th>
<th>SE</th>
<th>z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With - Baseline</strong></td>
<td>44% [26, 59]</td>
<td>2.96</td>
<td>0.58</td>
<td>5.13</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
### Table 8. Results of the frequentist and Bayesian analyses, Experiment 2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-value</th>
<th>Bonferroni-corrected p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regarding - With</td>
<td>5% [-9, 19]</td>
<td>0.05</td>
<td>0.31</td>
<td>0.18 &lt; .99</td>
</tr>
<tr>
<td>Which - Regarding</td>
<td>34% [17, 50]</td>
<td>2.13</td>
<td>0.44</td>
<td>4.80 &lt; .001</td>
</tr>
</tbody>
</table>

For the frequentist analysis, we show the estimate, standard-error, $t$-value and Bonferroni-corrected $p$-value of the tested contrast. For the Bayesian analysis we show the mean of the posterior distribution of each parameter of interest, together with a 95% credible interval (CrI).

#### 3.3 Discussion

Experiment 2 investigated to what extent speakers anticipate a dependency, namely a backward referring element, for different types of antecedents, via production. Results revealed a high proportion (>70%) of backward referring elements in all cases where a contextually relevant referent was available in the sentence fragment. This completion type was far more frequent than completion of other pronouns or lexical NPs. In addition, the baseline production rate of pronouns in the *if*-question condition was significantly lower than the BRE production rate in the other three conditions.

Despite the high proportion of BREs in all the three non-baseline conditions, *wh*-questions were significantly the most constraining, with an immediate gap appearing in almost all completions. We interpret this to reflect the effect of syntactic processes, which underlie the strong gap prediction. In contrast, the pragmatic constraint in the *regarding*-phrase condition did not seem to affect BRE production rates significantly more than the general contextual availability of an antecedent as in the *with*-phrase condition.

Since *wh*-questions revealed a higher proportion of BRE completions, the absence of an effect for the pragmatic constraint in the *regarding*-phrase condition cannot be due to a ceiling effect. Therefore, it seems that the discourse-prominence of the antecedent does not assert a strong influence in this task. One possible (if unlikely) explanation for this is that, for some reason, information structure is not considered in prediction. Another possibility is that
in this task, the parser tries to utilize every possible clue for generating the completion. If this is the case, a general context prediction of a BRE due to the availability of a referent (as in the with-phrase condition) is strengthened to match that of a discourse-prominent antecedent (in a regarding-phrase). The next experiment examines whether reading time measures of active dependency formation can better differentiate dependencies required by information structure considerations and those enabled generally by the context.

4. Experiment 3: Reanalysis costs
How does the processing difficulty associated with a failed gap prediction compare with other unrealized expectations? The third experiment examines the costs of disconfirmed expectations in online processing. For this purpose, we use the filled-gap design, in which a potential gap/BRE position is filled with a lexical NP. We expected the size of the "filled-gap effect", reflected in increased reading times on this NP, to vary as a function of dependency prediction strength, such that antecedents which entail a dependency more strongly would give rise to higher reanalysis costs.

4.1 Methods

4.1.1 Participants
Participants were 92 native speakers of Hebrew (according to self-report) (mean age: 25.37, range: 19-35). Seven participants were bilingual of Hebrew and either Russian or French, and the rest were monolingual. Participants gave informed consent and received payment (20 NIS, approximately 5 US dollars) for their participation.

4.1.2 Materials
Materials were based on sentences from the previous experiments, and included 24 sets, with four conditions each (see example sentences in Table 9). Experimental items followed the
filled-gap design, namely a lexical NP appeared in the direct object position of the first embedded verb, where a gap/co-referential argument would be predicted. The gap/pronoun was eventually realized within the clausal complement of that verb. Conditions included the three antecedent types (wh-question, regarding-phrase, and with-phrase) and an if-question baseline in which this additional referent did not appear, eliminating the possible prediction of a backward referring element.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sentence</th>
</tr>
</thead>
</table>
| Wh-question | ha-ovdim ša'alu evze kupai ha-menahelet ha-gvoha hixrixa et the-workers asked which cashier the-manager the-tall forced ACC

ha-šomeret ha-xadaša lesalek ba-šavua še-avar

the-guard the-new to.banish in.the-week previous

'The staff asked which cashier the tall manager forced the new security-guard to throw out last week.' |
| Regarding-phrase | ha-ovdim ša'alu legabev ha-kupai im ha-menahelet ha-gvoha the-workers asked regarding the-cashier whether the-manager the-tall

hixrixa et ha-šomeret ha-xadaša lesalek oto ba-šavua še-avar

forced ACC the-guard the-new to.banish him in.the-week previous

'The staff asked regarding the cashier if the tall manager forced the new security-guard to throw him out last week.' |
| With-phrase | ha-ovdim ša'alu et ha-kupai im ha-menahelet ha-gvoha

the-workers asked ACC the-cashier whether the-manager the-tall

hixrixa et ha-šomeret ha-xadaša lesalek oto ba-šavua še-avar

forced ACC the-guard the-new to.banish him in.the-week previous

'The staff asked the cashier if the tall manager forced the new security-guard to throw him out last week.' |
| If-question | ha-ovdim ša'alim ha-menahelet ha-gvoha hixrixa et ha-šomeret

the-workers asked whether the-manager the-tall forced ACC the-guard

ha-xadaša lesalek et ha-kupai ba-šavua šeavar

The staff asked if the tall manager forced the new security-guard to throw him out last week.' |
The staff asked if the tall manager forced **the new security-guard** to throw out the cashier last week.

Table 9. Example set from Experiment 3. The relevant antecedent is underlined; the critical NP is marked in bold; **ACC** = accusative case marker.

The 24 experimental items were distributed in a Latin square design across four lists. Within each list, the target sentences were combined with 40 grammatical filler sentences, for a total of 60 sentences. Since the experimental items all presented a dependency spanning a long distance, we included several types of dependencies among the filler sentences, to prevent adaptation to the materials of the experiment. Twenty filler sentences contained a short dependency, namely an embedded question with a gap or pronoun following the first verb (five sentences for each of the four structures in the experimental design). Fifteen other filler sentences contained neither a filler-gap dependency nor pronominal co-reference. These too mirrored the structures of the experimental items,10 with the exception of the **wh**-questions (which would be ungrammatical). The remaining five items were adjunct **wh**-questions.

### 4.1.3 Procedure

The experiment was built using IbexFarm (Drummond, 2018), and participants took part in the experiment online. Sentences were presented in a moving window, non-cumulative self-paced reading display: Sentences were initially masked by dashes with white spaces and

---

10 Note that in Experiment 1A, sentences containing a regarding-phrase with no subsequent pronoun received low acceptability ratings. For the current experiment, we constructed filler sentences of this type that would nonetheless be sensible and acceptable, by keeping the clause closely related to the regarding-phrase using strongly associated concepts like **apartment** and **lease** (e.g. "The real estate broker checked regarding the apartment whether the lease restricts animal housing").
punctuation intact; participants pressed the space bar to reveal each word. Before they began
the experiment, participants underwent a practice block of seven items. Seventy-five percent
of the sentences were followed by a ‘yes/no’ comprehension question.

4.1.4 Data Analysis
Participants were excluded from analysis if their performance on comprehension questions of
experimental items was not significantly different from chance level (resulting in the removal
of eleven participants), or if their average RT in experimental items was larger than 2.5 SDs
above the group’s average (resulting in the removal of three participants). For the remaining
78 participants, RTs higher than 2.5 SDs above the individual’s average RT were trimmed to
that cutoff (affecting 2.39% of the data). RTs shorter than 120ms were excluded (affecting
0.079% of the data). The accuracy rate of the remaining 78 participant was 82.76% on
experimental items, and 81.91% on filler items.

We analysed RTs from the critical (filled-gap) region, as well as from the spillover
word. We used a sliding contrast scheme of pairwise comparisons to detect an increase in RT
for every condition relative to the previous level. Namely, for every region, three
comparisons were made: with-phrase condition vs. if-baseline, regarding-phrase condition vs.
with-condition, and wh-question vs. regarding-phrase condition.

The data were analysed in R using mixed-effects linear models. We conducted two
types of analyses: a frequentist analysis (null hypothesis significance testing) and a Bayesian
analysis using weakly informative priors. For the frequentist analysis, the model was fitted
using restricted maximum likelihood (REML) and p-values were extracted by Satterthwaite
degrees of freedom approximation using the lmerTest package (Kuznetsova et al., 2017). We
corrected p-values using the Bonferroni correction for three comparisons. The Bayesian
hierarchical models were fit using Stan (Carpenter et al., 2017), via the brms package
(Bürkner, 2017). Modelling decisions for both analyses are described below. The code of
both analyses and the associated data are available at https://osf.io/9XB2Y/ as supplementary materials, along with a summary of an earlier (low-power) version this experiment.

Null hypothesis significance testing: To remove irrelevant effects, we used a two-step analysis. First, a basic model using log-transformed RTs was fitted to all the data, namely all the words in all the sentences, including both filler and experimental items (but excluding practice items). This model included word length, position of the word in sentence and trial number as fixed effect predictors, as well as a random intercept for participants. Residuals of this model were then entered into a second model. In this second stage, only data from critical words in the experimental items (filled-gap region / spillover word) were considered. This model included as fixed effects the experimental factors and the log RT of the previous word. A maximal random-effect structure was used in this model, with both intercepts and slopes. We fitted the second stage model separately for the critical word (the filled-gap noun) and the spillover region (the adjective following the critical NP).

Bayesian analysis: We fitted two separate hierarchical models, one for the critical word (the filled-gap noun), and one for the spillover region (the adjective following the critical NP). We used a LogNormal likelihood function to model the raw values in milliseconds (ms). The models included a full random effects structure and only the experimental fixed effect (i.e. we did not use the two-stage modelling procedure, for easier interpretation of the models’

11 logRT ~ experimental + rcs(position) + log(trial) + length + (1|subject)
The restricted cubic spline (RCS) function is used here for word position in order to capture non-linear effects of this factor (e.g. wrap-up effects).

12 ResidualRT ~ condition + previousWordRT + (1+ condition | subject) + (1+ condition | item)
Due to convergence failure the slope with the lowest variance was removed, resulting in the following model:

ResidualRT ~ condition + previousWordRT + (1+ condition | subject) + (1+regardingVSwith+whVSregarding | item)
results). In the text, we report the estimates along with their credible intervals (CrI) transformed back to milliseconds, for easier interpretation.

We used regularizing, mildly informative priors (Gelman et al., 2014): a standard normal distribution $N(0, 1)$, was used as the prior distribution for the fixed effects and for the standard deviation parameters. For the intercept, we had a wider prior of $N(0,10)$. Finally, for the correlation matrices of the random effects, priors were defined using the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009).

For each of the models, we sampled from the posterior distribution by running four MCMC chains at 2000 iterations each. The first half of the samples was discarded as warm-up samples. Convergence was checked using the R-hat convergence diagnostic and by visual inspection of the chains (Gelman et al., 2014).

**4.2 Results**

Mean reading times in the critical regions for the different conditions are presented in Table 10 and Figure 3a. Word-by-word reading times for the different experimental conditions are portrayed in Figure 3b. We report pairwise comparisons between every two consecutive conditions, in order to test the contribution of each type of dependency to the filled-gap effect. A retrospective power analysis of this experiment is available in the appendix.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Critical NP</th>
<th>Spillover region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wh-question</td>
<td>651 (19.9)</td>
<td>553 (12.7)</td>
</tr>
<tr>
<td>Regarding-phrase</td>
<td>565 (17.8)</td>
<td>559 (16)</td>
</tr>
<tr>
<td>With-phrase</td>
<td>555 (16.2)</td>
<td>508 (12.5)</td>
</tr>
<tr>
<td>If</td>
<td>565 (16.3)</td>
<td>528 (13.1)</td>
</tr>
</tbody>
</table>
Table 10. Mean (SE) RTs at the filled-gap NP and spillover by condition, Experiment 3.

Figure 3. A: Mean RTs by condition in the filled-gap and spillover word (Error bars mark +/- 1SE). B: Word-by-word mean RTs by condition. The critical area is shaded.

The workers/1 asked/2 regarding the cashier/3 whether/4 the manager/5 the-tall/6 forced/7 acc the-guard/8 the-new/o to.banish/9 him/11 in.the-week/12 previous/13

Frequentist analysis. The results are summarized in Table 11. On the critical NP, the analysis failed to reveal a significant difference between with-phrases and the baseline, or between regarding- and with-structures. However, a significant increase in RT was observed on the critical NP in wh-questions relative to regarding-phrase sentences ($t = 2.92, p = .02$).

At the spillover region, there was still no evidence for increased RTs in the with-condition, relative to baseline. However, we observed a significant increase in RTs in the regarding-condition relative to the with-condition ($t = 2.52, p = .047$). RTs in the wh-question condition did not differ significantly from those in regarding-condition.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$T$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical region:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With-phrase vs. if-baseline</td>
<td>0.016</td>
<td>0.026</td>
<td>0.61</td>
<td>&gt; .99</td>
</tr>
<tr>
<td>Regarding-phrase vs. with-phrase</td>
<td>0.004</td>
<td>0.027</td>
<td>0.15</td>
<td>&gt; .99</td>
</tr>
<tr>
<td>Wh-question vs. regarding-phrase</td>
<td>0.077</td>
<td>0.026</td>
<td>2.92</td>
<td>.02</td>
</tr>
</tbody>
</table>
Spillover region:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>With-phrase vs. if-baseline</td>
<td>0.001</td>
<td>0.021</td>
<td>0.035</td>
<td>&gt; .99</td>
</tr>
<tr>
<td>Regarding-phrase vs. with-phrase</td>
<td>0.063</td>
<td>0.025</td>
<td>2.52</td>
<td>.047</td>
</tr>
<tr>
<td>Wh-question vs. regarding-phrase</td>
<td>0.020</td>
<td>0.027</td>
<td>0.9</td>
<td>&gt; .99</td>
</tr>
</tbody>
</table>

Table 11. Results of the frequentist analysis: Estimate, standard-error, t-value and Bonferroni-corrected p-value of the tested contrasts.

Bayesian analysis. The Bayesian analysis produced similar results. The results of the model are summarized in Table 12 and Figure 4. The estimates and CrI of the posterior distributions suggest a filled-gap effect in the critical region in the wh-question condition: the mean of the posterior is 60 ms with CrI [29, 92] ms. In addition, we found support for a filled-gap effect in the spillover region in the regarding-condition: the mean of the posterior is 31 ms with CrI [6, 57] ms. The model did not show evidence for other differences between the conditions.

<table>
<thead>
<tr>
<th></th>
<th>Posterior mean and CrI (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical region:</td>
<td></td>
</tr>
<tr>
<td>With-phrase vs. if-baseline</td>
<td>-12 [-41,16]</td>
</tr>
<tr>
<td>Regarding-phrase vs. with-phrase</td>
<td>2 [-26, 29]</td>
</tr>
<tr>
<td>Wh-question vs. regarding-phrase</td>
<td>60 [29, 92]</td>
</tr>
<tr>
<td>Spillover region:</td>
<td></td>
</tr>
<tr>
<td>With-phrase vs. if-baseline</td>
<td>-14 [-37, 8]</td>
</tr>
<tr>
<td>Regarding-phrase vs. with-phrase</td>
<td>31 [6, 57]</td>
</tr>
<tr>
<td>Wh-question vs. regarding-phrase</td>
<td>9 [-19, 37]</td>
</tr>
</tbody>
</table>

Table 12. Posterior means of the tested contrasts, with 95% Bayesian credible intervals (CrI).

We further looked into effect size estimates (derived by the Bayesian analysis) for the two statistically robust effects: the filled gap effect detected for wh-questions (the contrast
between wh-questions and *regarding*-phrases, at the critical word) and that observed in the *regarding*-phrase condition (the contrast between *regarding*- and *with*-conditions, at the spillover region). We found equivocal evidence that the magnitude of the effect in the wh-questions was larger than in regarding-phrases. The effect in wh-questions produced a numerically bigger estimate than that in *regarding*-phrases (with a posterior mean of 60 vs. 31). However, the CrI of those contrasts overlap on one side (see figure 4). Therefore, we conclude that the analysis only weakly supports the premise of a magnitude difference between the two effects of interest.

![Figure 4. Posterior means and credible intervals (CrI) by condition, Experiment 3.](image)

**4.3 Discussion**

In Experiment 3, we tested the processing costs ensued by a failed prediction for a gap/BRE, reflected by the filled-gap effect, as indexing the strength of the prediction. We observed active dependency formation in *wh*-questions, indicated by a filled-gap effect at NP following the first sub-categorizing verb, above and beyond processing costs in this region in the *regarding*-phrase condition. The *regarding*-phrase condition seemed to exhibit a belated filled-gap effect, with increased RTs compared to the *with*-phrase condition at the spillover region. Finally, we failed to find evidence for a filled-gap effect in *with*-phrase sentences.
RTs in this condition did not significantly differ from those in the baseline if-condition (both at the critical NP and at the spillover region).

We suggest that the differences in the pattern of the filled-gap effects reflect different degrees of prediction associated with pragmatically- and syntactically-motivated dependencies. Information structure biases (in the regarding-phrase condition), create a structural preference for an early argument referring to that antecedent. This incites some prediction for a dependency, which is then reanalysed following the filled-gap NP. Yet, syntactic licensing pressures (in filler-gap dependencies) result in a faster and/or stronger prediction for a gap – as expressed in earlier (and possibly larger) reanalysis costs. Namely, predictive dependency formation arises for pragmatically motivated dependencies, but it does not fully account for predictive dependency formation in filler-gap dependencies. In section 4.3.1 below, we conjecture as to the different nature of the reanalysis in the two cases.

It should also be mentioned that in this experiment, we failed to find support for the maximal interpretation account of predictive dependency formation. The experiment did not detect a filled-gap effect following antecedents which are not syntactically or pragmatically prominent (i.e. in the with-phrase condition). Moreover, the filled-gap effect in the regarding-phrase condition was observed relative to the with-phrase condition, and exceeded any processing costs which may be associated with this structure. This suggests that predictive dependency formation is not fully reducible to anticipatory processes of general interpretation (see Kamide, 2008 for review of this kind of effects).

4.3.1 Possible accounts for the delayed reanalysis effect
Processing difficulty was observed immediately (on the filled-gap noun) in the wh-condition, but in the spillover position in the regarding-phrase condition. How can we make sense of the divergence in the position of reanalysis? The nature of the self-paced reading task makes it hard to establish the fine-grained time-course of reanalysis. Thus, the belated effect could
reflect a delay in different phases of the reanalysis process. One possibility is that in pragmatically-motivated dependencies (relative to syntactically-motivated ones), the parser is late to detect the required reanalysis. This might mean that readers allocate less attention to confirming the well-formedness of such dependencies (relative to filler-gap dependencies).

It is also possible that the late reanalysis effect in regarding-phrase sentences does not reflect a delay in detecting the need for reanalysis, but rather in advanced phases of the reanalysis. Namely, the delayed effect could indicate that readers more easily postpone full reanalysis in processing these pragmatically-motivated dependencies. This could suggest that reanalysis costs in this case are not too severe, so that readers still engage in parsing of the input while reanalysing, which makes the effect "spread out" to the following region.

In general, it is possible that the motivation for structure building affects the parser's commitment to it and thus to the costs of its reanalysis. Based on the results of this study, it is conceivable that lower commitment, in pragmatically-motivated dependencies, results in reallocation of some reanalysis costs to a later stage of processing. This may be similar to other factors which affect reanalysis difficulty. For instance, increased length of the ambiguous region can make reanalysis more costly in garden-path sentences (Ferreira & Henderson, 1991; Tabor & Hutchins, 2004). These “digging-in” effects are also understood as evidence that the parser becomes more committed to the wrong structural analysis.

4.3.2 Alternative accounts

One possible alternative to our account might attribute the results to possible differences in surprisal between the conditions. Surprisal theory (Hale, 2001; Levy, 2008) posits that the cost of processing a word in the input depends on that word’s conditional probability in the sentential contest (i.e. the probability of that word given the previous input). It could be suggested that the constructions we investigate are not comparable in terms of predictability
of the filled-gap element, and that this confound out results. Indeed, as we observed in Experiment 2, an NP in not produced at the same rate across the different conditions in our critical region (rate of NP production was 74% in the if-condition; 21% in the with-phrase condition, 14% in the regarding-phrase condition, and 2% the wh-question condition).

To test this hypothesis, we conducted an additional analysis of Experiment 3, using a surprisal measure derived from the production results. As 20 out of the 24 sets of Experiment 4 were based on materials from Experiment 2, we were able to derive by-item surprisal measures for most items in the four conditions. If surprisal drives the contrast between the conditions, we might expect that this measure would account for the data better than our experimental contrast. However, the results failed to exhibit a significant effect of surprisal both at the critical word (Estimate = 0.006; SE = 0.036; t = 0.17; p = .86) and at the spillover region (Estimate = 0.022; SE = 0.033; t = 0.66; p = .51), and the results of the experimental contrasts showed the same pattern of significance as in the main analysis.

Another possible account for our data might attribute the contrast between the wh-question and regarding-phrase conditions to a contrast between prediction of a null element (in filler-gap dependencies) and that of a pronoun (in the other conditions). It could be suggested that a gap is not predicted per se, but is assumed to be already in the input, and thus is fully integrated once the verb is processed. Pronouns, however, cannot be assumed to already be in the input, which might delay or weaken active dependency formation in the case

---

14 It should also be noted that frequency-based prediction and other more "principled" parsing strategies are inherently confounded in many studies of sentence processing. It is not possible to exclude the possibility that our data reflects the distribution of different structures in the language. Yet, if we reduce the effect to distribution patterns, we should also take into account that those distributions depend on linguistic processing (via production mechanisms). Therefore, we believe that moving the focus to frequency-based prediction eventually means pointing to underlying principles of production (e.g. MacDonald, 2013), that are not frequency-based.

15 Full specification and results of this model are available along with the code at https://osf.io/9XB2Y/.
of *regarding*-phrase antecedents. However, active dependency formation was recently reported to arise even when the predicted resolution of the dependency is lexically realized. In Hebrew, some dependencies can (or even must) be grammatically resolved with a resumptive pronoun. Keshev & Meltzer-Asscher (2017) observed a filled-gap effect, on the critical NP, in environments where only a resumptive pronoun is licensed. Thus, we assume the parser can commit to a dependency prediction to a similar extent in filler-gap and filler-pronoun dependencies. Yet, we recognize that comparing patterns from different experiments can be problematic, and that this distinction may have some effect on our data.

5. General discussion
The current study aimed to clarify the role of syntactic pressures, pragmatic considerations and general context availability in the prediction of dependencies. Previous studies investigating filler-gap dependency formation confounded the presence of syntactic motivation with the existence of an antecedent, and therefore could not determine whether active dependency resolution results from a structural principle or is part of general context-based prediction. The results obtained in the current set of experiments provide evidence that the "active filler" strategy may be partly motivated by discourse considerations, but is not fully reducible to them. The results suggest that active dependency formation is faster or more committing, in syntactically-motivated predictions (i.e. in filler-gap dependencies) relative to pragmatically-motivated ones.

Specifically, in this study we examined the prediction of long-distance dependencies with three different antecedent types: (i) fillers, which syntactically require co-reference; (ii) antecedents in *regarding*-phrases, which initiate a pragmatically-based prediction for a co-referring element; and (iii) antecedents which are contextually available but do not prompt any syntactic or discursive incentive for dependency formation (in *with*-phrases).
We observed evidence that the parser predicts an early resolution for filler-gap dependencies, in both production (Experiment 2) and comprehension (Experiment 3). In the case of discourse prominent antecedents (in regarding-phrases), we also found a substantial (though lower) rate of dependency prediction in production. This structures also exhibited a belated filled-gap effect. Taken together, the results suggest that predictive dependency formation occurs for both types of dependencies, but either on different timescales and/or with different degrees of commitment. Specifically, it could be hypothesized that in pragmatically-motivated dependencies some phases of the reanalysis process are postponed due to engagement in other processing procedures. From a methodological point of view, our results suggest that early filled-gap effects are better than spillover effects as indicators of syntactically motivated dependency formation.

General contextual predictions (represented in the with-phrase condition) exhibit yet another level in the gradient of dependency prediction degrees. While the production experiment revealed a strong preference for using co-referential arguments even in the with-phrase condition, we failed to detect a cost for disconfirmed predictions for this sentence type in Experiment 3. We believe there are two ways in which this seeming contradictory pattern of results can be settled: it could result from an inflated estimate of prediction processes in production tasks, or from difficulty in detecting such prediction in the filled-gap paradigm.\footnote{Another possible source for the contrast between Experiments 2 and 3 could be demographic. In particular, the mean age of participants was higher in Experiment 2. We do not know of any reason to suspect that older readers over-predict only in the general context conditions (with-phrase sentences). However, this possibility should be noted.}

Considering the first possibility, production tasks may bias participants to use any possible contextual cue, because using given information can facilitate production, removing the need to come up with an additional discourse referent and retrieve a new noun from the lexicon with little cues. In comprehension, on the other hand, participants are not required to
think up a referent by themselves, but merely need to recognize it in the text. Therefore, in production-based tasks, the strong preference for re-usage of referents might give rise to effects similar to those generated by a pragmatically-motivated prediction for co-reference.

Alternatively, it could be that the with-phrase condition does initiate some prediction processes, which were nevertheless not detected in the filled-gap experiment. There are several reasons which could be suggested for why such an effect would not be detected. First, it is possible that the filled-gap effect in this case is smaller, and that Experiment 3 did not have enough statistical power to detect this subtle difference (see the appendix for a power analysis of the observed effects). Second, it could be that the prediction is not generated fast enough, and thus affects only late measures like production rates (i.e. if the prediction is not made by the time the parser encounters the "filled-gap" NP, no reanalysis is required). Lastly, the context-based prediction in the with-phrase condition might generate a qualitatively different process. Low commitment of the parser to the prediction can mean, for example, that a conflict between this prediction and the actual input is not costly at all, in line with the distinction, from lexical prediction studies, between benefits for confirmed predictions and costs for disconfirmed predictions as measures of pre-activation (Van Petten & Luka, 2012). Namely, it could be that the filled-gap design is ill-formed for detecting certain kinds of predictions, since it is based on observing processing costs of failed predictions, rather

---

17 In lexical prediction an important distinction has been made between ease of integration due to partial overlapping activation, and predictive pre-activation of a constituent (Van Petten & Luka, 2012). Ease of integration is exhibited in comparing completions which are congruent or incongruent with a prediction (e.g. using the N400 effect). This is believed to indicate low level prediction (possibly via passive spreading activation from previous words in the context). If the prediction is strong and specific enough, the processing of a different completion becomes straining relative to cases (of similar plausibility) where no specific prediction was made (e.g. Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Ness & Meltzer-Asscher, 2017; Otten & van Berkum, 2008; Thornhill & Van Petten, 2012, among others). The filled-gap design is more similar to the latter type of finding, as it compares the consequences of failed prediction to a no-prediction baseline.
than processing benefits (i.e. ease of activation or integration) for correct predictions. Future research could look into this by testing contextual prediction (in structures like the with-phrase condition of this study) using other paradigms, which target integration ease rather than reanalysis costs.

To conclude, we propose that our results exhibit three different degrees of prediction: Contextually available antecedents may incite a preference for co-reference, yet this prediction is weak and can be detected only in some tasks; Pragmatic motivations (e.g. information structure considerations) can provoke predictive dependency formation whose disconfirmation is costly; Lastly, syntactic licensing pressures enhance the predictive process such that dependency formation is observed at a higher rate (in production) and in earlier measures (in comprehension).
Appendix: Retrospective power analysis for Experiment 3

We carried out a power analysis for Experiment 3 according to the procedure used by Vasishth, Mertzen, Jäger & Gelman (2019). We used a range of possible effect sizes, based on the estimate of the effect size extracted from the Bayesian analysis and its credible interval boundaries. We generated repeated data simulations to determine the proportion of cases where an effect of the relevant size is detected. We defined the discovery rate as the proportion of cases where the absolute t-value of the effect of interest exceeds 2. The code and the associated data are available at https://osf.io/9XB2Y/ as supplementary materials.

Table A1 summarizes the results of the power analyses for the two discussed effects: (i) the contrast between wh-questions and the regarding-phrase condition at the critical region; and (ii) the contrast between regarding and with conditions at the spillover region.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Low boundary of CI</th>
<th>Estimate</th>
<th>High boundary of CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical word: wh vs. regarding</td>
<td>29 ms</td>
<td>60 ms</td>
<td>92 ms</td>
</tr>
<tr>
<td></td>
<td>54%</td>
<td>96%</td>
<td>100%</td>
</tr>
<tr>
<td>Spillover region: regarding vs. with</td>
<td>6 ms</td>
<td>31 ms</td>
<td>57 ms</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>60%</td>
<td>98%</td>
</tr>
</tbody>
</table>

Table A1. Retrospective power analysis of Experiment 3. We report a range of effect sizes derived from the effect size estimate and its 95% credible intervals, and the corresponding power estimates (the probability of detecting the effect under repeated sampling, expressed as a percentage).
References


